DOCUMENT RESUME

ED 221 578 TM 820 607

AUTHOR Rubinstein, Sherry A.; And Others
TITLE The State of the Art of Teacher Certification Testing.
INSTITUTION National Evaluation Systems, Inc., Amherst, Mass.
PUB DATE Mar 82
NOTE 115p.; Papers presented at the Annual Meeting of the National Council on Measurement in Education (New York, NY, March 20-22, 1982).

EDRS PRICE MF01/PC05 Plus Postage.
DESCRIPTORS *Competency Based Teacher Education; Criterion Referenced Tests; Elementary School Teachers; *Minimum Competency Testing; Secondary School Teachers; *Standards; *Teacher Certification; Teacher Education Programs; *Teacher Evaluation; Teacher Qualifications

ABSTRACT

A series of five papers is presented. Sherry Rubinstein, and others, characterize the changes of teacher certification programs and reflect on the factors propelling and influencing the direction of those changes such as increased emphasis on the description and testing of the skills and knowledge of prospective teachers, and the adoption of criterion-referenced measures to assess teacher skills and knowledge. Katherine Vorwerk and William Gorth present a general model for developing the formal testing component of a certification program. The model includes: (1) developing certification requirements; (2) deciding how to assess requirements; (3) defining measurement strategies and instruments; (4) handling logistical issues of assessment; and, (5) communicating and using assessment results. Michael Priestley explores various approaches to assessment for initial teacher certification. Conceptual issues are considered in relation to test design, assessment for entry to a teacher education program, exit credentialing, and classroom performance assessment. Paula Nassif reviews technical issues of teacher certification testing, focusing on standard setting and equating, and validity and job analysis. Scott Elliot presents current applications of job analysis methodology to teacher certification testing. (PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED221578

THE STATE OF THE ART OF TEACHER CERTIFICATION TESTING

National Evaluation Systems, Inc.

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

P. M. Nassif

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 820 607

THE STATE OF THE ART OF TEACHER CERTIFICATION TESTING

Paula M. Nassif, Organizer

Lester M. Solomon, Moderator
Georgia Department of Education

The Changing Nature of Teacher Certification Programs

Sherry A. Rubinstein
Matthew W. McDonough
Richard G. Allan

National Evaluation Systems, Inc.
Amherst, MA

Common Themes in Teacher Cerftification Testing Program Development and Implementation

Katherine E. Vorwerk
William Phillip Gorth

National Evaluation Systems, Inc.
Amherst, MA

Variations in Approaches to Assessment

Michael Priestley

National Evaluation Systems, Inc.
Amherst, MA

Teacher Certification Testing Technical Challenges: Parts I & II

Paula M. Nassif
Scott M. Elliot

National Evaluation Systems, Inc.
Amherst, MA

Symposium presented at the Annual Meeting of The National Council on Measurement in Education, New York, 1982

THE CHANGING NATURE OF TEACHER CERTIFICATION PROGRAMS

Sherry A. Rubinstein

Matthew W. McDonough

Richard G. Allan

National Evaluation Systems, Inc
30 Gatehouse Road
Amhest, MA 01004

A Paper Presented at the Annual Meeting
of The National Council on Measurement
in Education, New York, 1982

Introduction

Early in the nineteenth century, the sole credential required of teachers of public school children was basic proficiency in reading, writing, and arithmetic. With the advent of mass compulsory education later in the century came the states' interests in extending these criteria to include proficiency in professional techniques and specific subject-matter knowledge. These three aspects--basic skills, competence in teaching techniques, and knowledge of subject matter to be taught--have continued through to the present as the mainstays of teacher assessment systems.

This characterization seems to suggest and underscore a considerable consensus about and continuity over time in the important aspects of teacher evaluation--although some would rather interpret this status as a reflection of the slow growth in our understanding of the elements of effective teaching and how to test for their presence. Naysayers notwithstanding, the last decade has been marked by dramatic change in approaches to credentialing public school teachers. The change has been not so much in the primary domains of competence subjected to scrutiny, but in the degree of emphasis accorded them, the manner in which they are characterized, and the manner in which they are assessed.

The nature of the change in credentialing practice is evidenced by the significant nationwide increase in efforts to reexamine and modify those state-level programs charged with the responsibility of licensing teachers. Licensure is the "process by which an agency of the government grants permission to an individual to engage in a given occupation upon finding that

the applicant has attained the minimal degree of competency required to ensure that the public health, safety, and welfare will be reasonably well protected" (U.S. Department of Health, Education, and Welfare, 1977, p. 4). An individual without a teaching license from a particular state is legally barred from the practice of public school teaching in that state. The closely related process of certification grants the use of a title (e.g., "teacher") to an individual who has met a predetermined set of standards or qualifications set by a credentialing agency (Shimburg, 1981). This distinction between licensure and certification having been made for the record; the commonly used generic referent "teacher certification program" hereinafter will denote individual state government policies and procedures regarding the granting of teacher licenses.

Prior to the late 1960s, most states credentialed prospective teachers on the basis of successful completion of a teacher education program of study. Only some states went so far as to require accreditation or "approval" of such programs; and only some states took the additional measure of requiring entrants into the teaching field to pass a nationally standardized, norm-referenced test. Such state policies had been stable for a considerable length of time, which suggested a prevailing opinion that certification programs were fulfilling their purpose. From the lack of controversy, one could conclude that most groups and individuals concerned with public education were satisfied that these programs were adequate to ensure that unqualified individuals were excluded from teaching and that all qualified applicants had fair and unbiased access to the profession.

The decade of the 1970s stands in marked contrast. During this time, teacher certification programs were taken to task by a variety of interest groups concerned with the quality of teaching in the nation's schools, and state departments of education faced strong and often contradictory demands for change. As a result, teacher certification programs were subjected to considerable scrutiny and underwent extensive changes. The purpose of this paper is to characterize these changes--particularly those related to tests and measures--and to reflect on the factors that perhaps propelled and certainly influenced the direction of those changes. In doing so, the authors will first call upon empirical evidence to document the existence and extent of the change observed and argue that the significant features of the change are (a) new and different emphases in the description and testing of the skills and knowledge which prospective teachers should possess; and (b) increasing adoption of criterion-referenced measures to assess the skills and knowledge so described. These changes will then be analyzed in terms of their relationship to events and factors in three separate spheres: (1) the general political environment, (2) the legal/regulatory environment, and (3) the educational/research environment. In summary, the authors will conclude that the changes are having or will have a variety of positive effects.

Evidence of Change

Substantiating claims of change in teacher certification programs is not a difficult task. That change "in the air" was evident and was publicized as early as 1975 when a study by Pittman (1975) revealed that between 1970 and

1975 every state in the Union had considered the idea of modifying teacher certification practices to incorporate the then-new principles of competency-based education. This spate of activity took a variety of forms including the appointment of study panels, the commissioning of position papers, the hosting of conferences, and the review of concrete proposals. These activities at a minimum suggested an interest in re-analyzing teacher certification requirements and, in a significant number of cases, this interest was followed by action. A number of states made significant modifications to their existing certification programs; others chose to design totally new programs to replace existing ones. Changes were variously brought to bear on the policies and practices of all four phases of teacher certification programs, those effective (1) upon admission to teacher training programs; (2) upon completion of such a program (initial certification); (3) during the first year of incumbency in a teaching position; and (4) during later incumbency (certification renewal).

One major form of revision affected the common policy that automatically granted certification to a graduate of any teacher education program. During the period of 1970 to 1975, 26 states revised such a policy and implemented a system of "approving" teacher education programs (Pittman, 1975). By far the most dramatic action (or at least the most publicly visible one), however, was to require that graduates of teacher education programs pass a state-sponsored test to obtain a license to teach. Between 1977 and 1981, 16 states enacted legislation or state board of education policy that either changed or initiated tests whose purpose was state licensing of teachers. Table 1 presents a list of the 16 states making substantive changes in one or more testing components of their certification programs between 1977 and 1981 and describes the

TABLE 1

Cross-State Matrix of Program Elements as of January 1982

	Entry to Teacher Ed. Program			Initial Certification					Further Certification Requirements		Renewal Requirements	
	Required GPA	Entrance Exam	Interview	Completion of Approved Program	Student Teaching	Basic Skills Test	Professional Studies Test	Content Area Tests	Internship	Evaluation	Teaching Experience	Semester Hours
ON-GOING												
Florida		x	x	x	x	CRT	CRT		x	x		x
Georgia	x	x	x	x	x			CRT	x	x		x
Louisiana	x	x	x	x	x			VNTE				
Mississippi		x	x	x	x	NTE	NTE	NTE				
Virginia			x	x	x	VNTE	VNTE	VNTE				
W. Virginia	x		x	x	x	*	*	*		x	x	x
IMPLEMENTING 1981-1982												
Alabama		x	x	x	x		CRT	CRT	x	x	x	
Arizona			x	x	x	CRT	CRT					
Arkansas			x	x	x	VNTE	VNTE	VNTE				
No. Carolina		x	x	x	x	VNTE	VNTE		x			x
Oklahoma	x	x	x	x	x			CRT	x	x		x
So. Carolina		x	x	x	x			VNTE/ CRT	x	x		x
Tennessee		x	x	x	x			NTE				
IMPLEMENTATION 1982-1986												
California			x	x	x	TBD		TBD			x	x
New York			x	x	x	TBD	TBD	TBD				
Texas		TBD	x	x	x	TBD		TBD				

CRT--Criterion-Referenced Test
NTE--National Teachers Exam (not locally validated)
VNTE--National Teachers Exam (locally validated)
TBD--Form of Test to be Determined

* Board motion pending to develop CRT instruments

discrete elements of these programs within all four phases of the certification program. Six of these states have since implemented the changes, while in seven others the changes become effective this year; in the other three states refinement and planning are in progress for implementation in the next few years.

The program elements depicted in Table 1 represent increased rigor in the entire teacher certification process. The move toward more widespread adoption of the approved-program model reflected the imposition of more stringent requirements in an effort to upgrade programs and to improve the quality of the professionals they graduated. All of the 16 states represented now have an approved-program requirement. Noteworthy activity is occurring in at least two other states. The New Jersey State Board of Education is considering imposing more stringent requirements on the curricula of teachers' colleges, and Connecticut is involved in related deliberations on ways to improve teacher education.

Nature of Testing-related Changes

More significant for present purposes are those requirements that involve changes in testing practices: (a) testing of prospective program entrants, and (b) testing of program graduates as eligible and prospective license holders. An example of the former is Alabama's newlyinstalled English Language Proficiency Test which assesses basic skills in reading, writing, language skills, and listening. It is the installation of tests such as this one that

reveal a heightened emphasis on "the basics" in the screening of prospective teachers. This trend is mirrored in end-of-program testing. An increasing number of states are including a basic skills test as one component of initial certification requirements; Florida's new program is a prime example.

That in more and more states graduates must pass a state-mandated test over and above fulfilling all other course and program requirements is itself evidence of increased stringency in certification programs. This evidence is less compelling, however, than the changing character of the tests being used. As Table 1 indicates, the most common tests in use are a nationally standardized norm-referenced test (the National Teacher Examination -the NTE), a locally validated norm-referenced test (the NTE subjected to a within-state validation process), and a customized criterion-referenced test (CRT). It is only in the last several years that the latter CRTs have come into common use for end-of-program testing. This trend has been concomitant with increasing specificity in the description of the skills and knowledge which entering teachers should possess, specificity characteristic of objective-referenced assessment.

Another significant feature of the change in initial certification testing is an increased emphasis on content-oriented tests. While some states have traditionally used the NTE Specialty Area Examinations, more and more states are funding the development of criterion-referenced tests in these and other areas. South Carolina's recent legislation, for example, called for customized development of CRT in eight teaching areas not covered by the NTE (including Trades and Industries, Distributive Education, German, Latin, Earth Science, Psychology, Speech and Drama, and Health). Georgia now has a total of 18

teaching field CRTs assessing prospective teachers' knowledge of the content to be taught to students in a variety of subject areas (including Agriculture, Music, Early Childhood, Middle Childhood, Communicative Arts, Business, Home Economics, Industrial Arts, French, and Spanish), with a nineteenth field (Health) currently under development. Oklahoma's new program being installed this year is by far the most extensively CRT based with 62 separate teaching area tests, including Journalism, Driver and Safety Education, an umbrella and subarea exams in Science (e.g, Zoology), Social Studies (e.g., Economics, Oklahoma History), Business Education (e.g., Accounting, Shorthand), and Language Arts (e.g., World Literature).

Even the foregoing recitation, however, underplays the range of content areas being assessed by CRTs. Special education is also receiving considerable attention. South Carolina has four separate special education area exams, Georgia has three, and Oklahoma has seven (counting the umbrella exam). There are also tests for other pupil personnel service positions. Oklahoma has seven: Psychologist, School Counselor, Speech Pathology, Psychometrist, Reading Specialist, Audiovisual Specialist, and Librarian. South Carolina has one (Speech Correction) and Georgia has two (Library Media and School Counselor) with three others currently under development. (There are also CRT certification exams for administrators -- Georgia's Administration and Supervision test and Oklahoma's three separate tests for superintendents, elementary and secondary school principals, respectively.)

The development and installation of these tests are strong indications of the increasing emphasis on content area (subject-matter) tests and the increasing adoption of criterion-referenced approaches to measurement. Other

changes have come hand-in-hand with these. The developmental process for teacher certification tests has been increasingly characterized by a strong validation effort. Examples are the local validation process to which the standardized NTE is being subjected in some states (see "VNTE" states on Table 1) and the full-scale job analyses which, as an early step in the development of CRTs, serve to identify the knowledge and skills viewed by job incumbents (teachers of the specific subject matter) as frequently used and important in their work (e.g., in Georgia and Oklahoma).

There is little doubt that recent developments in the nature and types of tests in use represent a significant change in teacher certification policy. These trends, however, did not develop in a vacuum. They have their sources in, or at least were influenced by, factors in three other areas: the general political environment, the legal/regulatory environment, and the education/measurement environment. Each of these areas is analyzed in the following sections in an attempt to unsort and identify factors postulated to bear a relationship to the changing nature of teacher certification programs.

The General Political Environment

The concept of "political environment" is here intended to denote the set of factors which, when taken together, constitute the sociopsychological and socioeconomic fabric of our collective lives. Thus, we distinguish from all other factors those which appear to be out of the purview or control of any single individual, group, agency, or institution. The indicators of the general political environment are readily perceptible and, over the past decade, one of the most obvious was an alarmingly pervasive dissatisfaction

with the outcomes of public education. This dissatisfaction, voiced and also fueled by the national media, included educators' frustration with a ten-year decline in SAT scores, parents' reports of functionally illiterate high school graduates, and business leaders' complaints about the lack of even minimally qualified entrants into the work force.

In the early 1970s, parents and other critics alike began demanding a "return to basics" as a means of assuring the accountability of local school systems. "Accountability" itself became a byword, if not a bona fide movement, and it targeted all tangible features and products of the schools. First, the spotlight was turned to students themselves; public pressure led legislatures and state departments of education, through the 1970s, to institute minimum competency test programs. These programs, while diverse in design, had the common purpose of reflecting the school systems' success or failure at teaching certain predefined "basics" to each and every student. These programs imposed consequences on students for a failure to perform at or above "minimally acceptable" levels.

An equally harsh light was cast on school curricula, including a look not only at the traditional "Three Rs", but at social studies, science, and a host of other subjects. Public pressure was exerted to increase the utility of what was taught to students, a continuation of the demand for relevance heard earlier in the 1960s. In response, educators began modifying curricula in form and/or substance, to focus on skills and knowledge useful to students in their economic, political, and social lives. The emphasis moved from what students should know to what students should be able to do, the latter being more observable and therefore more productive of answers to questions of accountability.

Throughout the decade, the mass media and popular press devoted substantial coverage to the "crisis in education" and the ability of the system to educate the nation's youth. It should have come as no surprise, then, that the focus broadened from an examination of the curriculum and student outcomes to include an appraisal of the agents of instruction: teachers themselves. From books such as Morris Kline's Why Johnny Can't Read (Kline, 1973) to a New York Times editorial (Montgomery, 1979) to a cover story for Time Magazine (Help! Teacher Can't Teach; 1980), the competence and ability of those who teach came under increasing attack. The public demanded assurances that teachers were qualified to do their jobs--to such an extent that it was estimated that the teacher testing movement, the most visible of all certification-related activities, was supported by 85% of U.S. adults (Foote, 1980).

It is noteworthy that the 1970s were characterized by these demands for accountability. The underlying factor might be isolated as a common preoccupation with economic pressures. The decade was beset by rapid inflation and diminishing resources which resulted in turning the public's attention away from perceived "luxuries" in education and spawned this new "back to basics" movement. It could be argued that, as an extension of concern about personal budgetary constraints, the consumers of education were asking (and continue to ask) what value they were getting for their education tax dollars. In the face of strong countervailing efforts by teachers' unions to "protect" incumbent teachers, the states' response to these consumer demands focused heavily on credentialing of prospective teachers. In many cases, the response was the very visible one of expanding and strengthening the initial certification testing components of their programs.

The Legal/Regulatory Environment

As public pressure was brought to bear on teacher certification programs, a number of legal and regulatory precedents were being set which influenced the direction of the movement. These were an outgrowth of Title VII of the Civil Rights Act and the Equal Employment Opportunity Commission (EEOC) Guidelines on Employee Selection Procedures. Additionally, there was the influence exerted by development of the 1974 version of the Standards for Educational and Psychological Tests (APA, AERA, NCME, 1974). The promulgation of these regulations and standards reflected increasing legislative, judicial, and professional concern with fair employment practices both in and out of education.

Legislation, regulations, and the Courts. Stated simply, Title VII of the Civil Rights Act of 1964 outlawed employment discrimination on the basis of sex, race, color, religion, or national origin and empowered the EEOC to enforce the stipulations of the law. The 1970 EEOC Guidelines, a revision of the first version published in 1966, included a set of stipulations founded on the premise that standardization and proper validation in employee selection procedures would build a foundation for the nondiscriminatory personnel practices required by Title VII. These stipulations (EEOC, 1970) included the following:

(a) empirical data should be made available to establish the predictive validity of a test, that is, the significant correlation of test performance with job-relevant work behaviors; such data must be collected according to generally accepted procedures for establishing criterion-related validity;

(b) where predictive validity is not feasible, evidence of content validity (in the case of job knowledge or proficiency tests) may suffice as long as appropriate information relating test content to job requirements is supplied;

(c) where validity cannot otherwise be established, evidence of a test's validity can be claimed on the basis of validation in other organizations as long as the jobs are shown to be comparable and there are no major differences in context or sample composition;

(d) differential failure rates (with consequent adverse effects on hiring) for members of groups protected by Title VII constitute discrimination unless the test has proven valid (as defined above) and alternative procedures for selection are not available; and

(e) differential failure rates must have a job-relevant basis and, where possible, data on such rates must be reported separately for minority and nonminority groups.

As a result of Title VII and the EEOC Guidelines, many concepts which had previously been the purview of psychometricians took on important legal ramifications. In the first major challenge to employment tests (Griggs v. Duke Power Company, 1971), the Supreme Court unanimously interpreted Title VII as prohibiting "not only overt discrimination but also practices that are fair in form, but discriminatory in operation" (p. 431). This decision decreed that absence of intent to discriminate was insufficient to justify the use of a test which had a disproportionate impact on protected minorities; even the employer with the best of intentions bore the responsibility of demonstrating "that any given requirement...(bears) a manifest relationship to the employment in question" (p.431). The Court further commented that the tenets of the Guidelines were "entitled to great deference" (p. 434) because they were drafted by the enforcing agency for Title VII. It was in this way that the concepts of "job relatedness" came to be incorporated into the law of employment testing (Bersoff, 1981) and virtually came to have the effect of law (Rebell, 1976).

Two other early cases are worthy of note. In Chance v. Board of Examiners (1972), the New York licensing exams for principals and other administrators were declared invalid for lack of job relevance. Later, in Albemarle Paper Company v. Moody (1975), the Court invoked EEOC and, in effect, established criteria to be used in proving whether employers' tests were job related. Specifically, the Court made reference to the importance of analyzing "the attributes of, or the particular skills needed in," (p. 432) a given job as a basis for creating a job-relevant test.

Most significantly for teacher certification programs was passage of a 1972 amendment (Public Law 92-261) to the Civil Rights Act which struck out the exemption for educational personnel in public institutions, extending the provisions of EEOC beyond private industry to state and local government agencies. Prior to the amendment, court challenges against public employers (e.g., Chance v. Board of Examiners) were initially brought on equal protection grounds under the Fourteenth Amendment which required only that employers demonstrate a rational basis for use of a test. Arguments only indirectly cited, but amassed consensual support for, EEOC Guidelines which were not technically binding at the time (Rebell, 1976). The 1972 Amendment paved the way for later litigation (e.g., United States v. State of North Carolina, 1975) which successfully challenged the NTE as a teacher selection test. For an excellent review of these cases and an overview of the law and teacher certification, see Licensing and Accreditation in Education: The Law and the State Interest (Levitov, 1976).

Throughout the decade, the concepts contained in the 1970 EEOC Guidelines were refined through the process of litigation and resulting Court opinion. Concurrently, various federal agencies were debating related issues, a debate which culminated in publication of the 1978 Uniform Guidelines (EEOC, CSC, Department of Labor, and Department of Justice, 1978), a document which contained "specific statements in most sections, in contrast to the more general statements of the 1970 Guidelines" (Novick, 1981, p. 1040). The intent was made clear: that a test must be a representative measure of the actual domain of skills used on the job and must be validated for its intended purpose.

Professional standards. A discussion of the regulatory environment affecting teacher certification testing cannot exclude the process whereby professionals and practitioners regulate themselves. An example of this self-regulation is reflected in the publication of the Standards for Educational and Psychological Tests (APA, AERA, NCME, 1974). Unlike earlier documents of its kind which stressed the obligations of test producers, the 1974 Standards addressed competency in testing practice and test use (Novick, 1981). Novick (1981) presents an excellent review of the evolution in professional standards over the last three-quarters of a century, but most revealing is his comment that this first document on test use "might not have happened, had it not been for the emergence of the social questions to which the EEOC Guidelines clearly responded, and the concomitant civil rights pressure of numerous advocacy groups" (p.1043).

The Standards display many similarities to the EEOC Guidelines and, in fact, both the 1974 document and its 1966 precursor were cited in numerous court cases (e.g., Albemarle) to bolster the credibility and importance of the Guidelines themselves (Bersoff, 1981). Beyond the emphasis on validation strategies, however, the Standards stressed the requirement to investigate potential bias in the measures and to report results for separate subsamples i.e., minority groups). Further, the Standards specified that any pass-fail scores used should be accompanied by "a rationale, justification, or explanation" (p. 66) for their adoption. It was provisions such as these which were taken seriously by the designers and implementers of the newer teacher certification program.

The combined impact. Taken together, Title VII, the EEOC Guidelines, resulting court challenges, and the Standards can be seen as catalysts and guides to the restructuring of teacher certification programs. Their impact is evidenced in several aspects of these programs:

(a) Because it has not been feasible to conduct predictive validity studies (based primarily on difficulties in obtaining reliable and valid measures of the criterion), the response has been to more fully incorporate other validation efforts. Increased attention is being paid to the validity of certification tests, and it is focused almost exclusively on content validity.

(b) The focus on content validity has greatly expanded the involvement of incumbent teachers and subject-matter specialists in the test development process, both through committee review work and participation in full-scale job analyses. It is through these methods that the test development process attends to the specific attributes of a job and provides evidence of the test's relevance to the job to which it applies.

(c) There is increased awareness of the potential for differential impact, with expanded efforts to include diverse interest groups in the test development process and to report test results separately for relevant minority groups.

(d) Finally, there has a been a shift toward the use of criterion-referenced, as opposed to norm-referenced, models of standard setting; a variety of methods incorporating expert judgments about the test items themselves are coming into more popular use.

These trends reflect the significant impact of the legal/regulatory environment on the design of teacher certification programs.

The Education/Measurement Environment.

It is in this final context, the education/measurement environment, that discussion focuses on factors within the purview of educators and psychometricians, rather than on factors external to the domain of education. Two distinct themes are to be examined: (a) theory development in relation to teacher education practice, specifically the growth of competency-based teacher education (CBTE), and (b) advances in measurement theory and statistical techniques relevant to criterion-referenced tests.

CBTE. The early 1970s saw the start of the CBTE movement, a newly conceived pedagogy for teacher education programs based initially on the already-established concept of mastery learning. Among 14 defining and ancillary features of CBTE, Hall & Houston (1981) included six which bear at least a surface relationship to the characteristics of the newer teacher certification programs:

(a) instruction focused on learner outcomes rather than on time in attendance;

(b) a priori description of the intended learner outcomes;

(c) introduction of subcompetency and competency statements;

(d) emphasis on mastery, at least to some minimum level of identified learning;

(e) de-emphasis on how well a student performs relative to other students in favor of emphasis on demonstration of desired outcomes; and

(f) clear and public communication of minimum levels of success with continual feedback on performance.

Even in its early days, there were optimistic predictions that CBTE would result in "new measures of teacher behavior" and "new criteria for certification" (Hall & Houston, 1981, p. 20). It was the basic tenet that instruction be objective based which was most influential. In the spread of CBTE to teacher training institutions, the pedagogy was rarely fully understood or fully adopted, but even where it was only superficially incorporated into an ongoing program, it included a focus on establishing objectives for learning. The debate surrounding CBTE therefore included in-depth examination and discussion of which skills and competencies teachers needed to develop. One chief product of this debate was the development of performance-based standards against which teacher competency could be judged. It was thus that CBTE provided the testing movement with the criteria necessary to develop clear, valid, job-relevant certification tests.

Tests, measures, and statistics. As CBTE provided the criteria to be measured (or at least fashioned the willingness to do so), it devolved to the measurement community to respond with appropriate tools and instruments. It became clear that the existing standardized norm-referenced tests could not fulfill the demand for content validity and tailored job relevance, for specification of objectives, or for scoring in comparison to preset criteria rather than in terms of group norms. Thus, the rapid growth in the demand for and use of criterion-referenced tests went hand-in-hand with the CBTE movement.

While it is beyond the scope of this paper to provide technical details, it is clear that the growth of CBTE- and CRT-supported (and, in turn, continued to be supported by) research and development of new measurement techniques. We have witnessed refinements in methods of defining domains (Popham, 1980) and generating statements of learning objectives (Popham, 1978), strategies for developing test items (Hambleton & Eignor, in press), and methods of setting cut scores (Nassif, 1978; Hambleton, 1980). There have also been significant advances in CRT-relevant statistics, including indices of reliability (Subkoviak, 1980), application of latent trait models (Cook, Eignor & Hutten, 1979), new approaches to item analysis (Berk, 1980), and new methods of investigating test item bias (Merz & Grossen, 1979). These technical developments went a long way toward enabling increased rigor in criterion-referenced testing conducted for public policy reasons. And, given the interface of researchers and practitioners, the need for a stringent, fair, and legally defensible system for certifying teachers fueled support for continuing technical refinements.

Summary and Conclusions

Early in this paper, the authors suggested a "bandwagon" effect in the increasing adoption of CRT-based teacher certification programs. In doing so, the intent was not to suggest automatically that "the band is playing the right tune," although the many CRT supporters in the professional community would like to think so. Yet, it can be argued that the recent trends toward increasing rigor in the teacher certification process is associated with a variety of positive effects:

(a) The visible nature of the change has increased the involvement of educators and special interest groups in debate over what teachers should know. This debate helps to fend off potential complacency that might thwart growth in our knowledge base about the constitutive elements of effective teaching.

(b) The movement has substantially increased communication about what the tests measure, a trend which serves to enhance the meaningfulness of test scores. This may be contrasted with the traditional scoring of NRTs which diverted attention away from test content in favor of person-to-group comparisons.

(c) The objectives-based construction of the tests enables test takers to learn, in advance, the expectations set for them, a condition which most recent research suggests contributes to maximizing performance.

(d) Completing the communication cycle, the newer certification programs entail expanded feedback to examinees on their performance, including indications of strengths and weaknesses with regard to specified domains on the tests.

(e) The objectives-based approach has also increased the utility of feedback to institutions about the performance of their graduates. The optimists among us (Hall & Houston, 1981) anticipate that, once the competency tests are installed, "teacher education programs will start preparing their students to a sufficient level of mastery of each test criterion" (p. 25). In essence this would constitute the upgrading of teacher education programs that was the initial intention of CBTE.

(f) The new legal implications have heightened the focus on incorporating the most state-of-the-art techniques in the measurement of the competencies of prospective teachers. The increased attention to technical rigor can only serve to further protect the test takers.

(g) Lastly, the visibility of these developments has turned a spotlight on the importance of the role of the public school teacher in American society. The controversies surrounding teacher certification testing have increased the outreach efforts of state departments of education to explain (or justify) their policies and practices. These efforts have, at a minimum, increased information sharing and the public's awareness of state efforts to fulfill accountability demands.

Notwithstanding these positive effects, there are several implications of the testing movement which deserve serious study. The first is a concern about the immediate teacher supply. With more stringent criteria for certification, fewer prospective teachers are likely to receive licenses, and school systems are likely to find it increasingly difficult to staff certain positions. Even under the hopeful assumption that, in response, teaching institutions will upgrade the skills of their graduates, there is little doubt that a significant time lapse will exist. In the meantime, state departments of education are likely to experience substantial pressure to implement politically expedient solutions to this problem.

Second, the reporting of test results for examinees on an institution-by-institution basis has already begun to engender political pressure to "reward or punish" institutions on the basis of their "performance." Where failure rates are excessive, for example, threats of loss of accreditation are not likely to be uncommon. In the face of these pressures, it will be increasingly difficult to ward off simplistic solutions to complex problems.

Third, and finally, the differential passing rates being observed for minority groups have direct implications for the proportion of minority group teachers in the nation's schools. While the testing programs and the tests themselves may be held to be valid on the basis of evidence they present, the implications of their use must be considered from the larger cultural and sociological perspective. Issues such as these are raised as caveats to the testing practitioner, in the interest of emphasizing that testing for public policy purposes must be conceived and implemented in a manner that is both professionally and socially responsible.

References

Albermarle Paper Co. v. Moody, 95 S CT. 2362 (1975).

American Psychological Association, American Educational Association, and National Council on Measurement in Education. Standards for educational and psychological tests and manuals. Washington, D.C.: American Psychological Association, 1974.

Berk, R. Criterion-referenced measurement: The state of the art. Baltimore: Johns Hopkins University Press, 1980.

Bersoff, D. Testing and the Law. American Psychologist, 1981, 36, 1047-1056.

Chance v. Board of Examiners, F. Supp. 203 (S.D. N.Y., 1971), Aff'd 458 F.2d 1167 (2D Cir., 1972).

Cook, L. L., Eignor, D. R., and Hutten, L. R. Considerations in the application of latent trait theory to objectives-based criterion-referenced tests. Laboratory of psychometric and evaluative research report. Amherst, MA: University of Massachusetts, School of Education, 1979.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice. Adoption by four agencies of uniform guidelines on employer selection procedures. Federal Register, 1978, 43, 38290-38315.

Griggs v. Duke Power Company, 401 U.S. 424 (1971).

Hall, G., Houston, R. Competency-based teacher education: Where is it now? New York University Quarterly, 1981, 3, 20-28.

Hambleton, R. K. Test score validity and standard-setting methods. In R. D. Berk (ed.) Criterion-referenced measurement: The State of the art. Baltimore, MD: Johns Hopkins University Press, 1980.

Hambleton, R., Eignor, D. A practitioner's guide to criterion-referenced test development, validation, and usage. Laboratory of psychometric and evaluative research report No. 70. Amherst, MA: University of Massachusetts, School of Education, 1979 (2nd ed.)

Help! teacher can't teach! Time Magazine, June 16, 1980.

Kline, M. Why Johnny can't read. New York: St. Martin's Press, 1973.

Levitov, B. Licensing and accreditation in education: The law and the state interest. Lincoln, Nebraska: University of Nebraska, 1976.

Merz, W. R., Grossen, N.E. An empirical investigation of six methods for examining test item bias (Final report grant NIE - 6-78-0067). Sacramento, CA: Foundation of California University, Sacramento, CA, 1979.

Montgomery, J. Can "teach" teach? New York Times, May 14, 1979.

Nassif, P.M. Standard-setting for criterion-referenced teacher licensing tests. Paper presented at the annual meeting of the National Council on Measurement in Education: Toronto, March, 1978.

Novick, M. Federal guidelines and professional standards. American Psychologist, 1981, 36, 1036-1047.

Pittman, J. Actions taken by state departments of education in developing CBTE certification systems. Paper delivered at the Association of Teacher Educators Annual Conference, New Orleans, February, 1975.

Popham, J. Criterion-referenced measurement. Englewood Cliffs, New Jersey: Prentice-Hall, 1978.

Rebell, M. The Law, the courts, and teacher credentialing reform. In B. Levitov (Ed.) Licensing and accreditation in education: The law and the state interest. Lincoln, Nebraska: University of Nebraska, 1976.

Shimberg, B. Testing for licensing and certification. American Psychologist, 1981, 36, 1138-1146.

Subkoviak, M. J. Decision-consistency approaches. In R. D. Berk (Ed.) Criterion-referenced measurement: The state of the art. Baltimore, MD: Johns Hopkins University Press, 1980.

U.S. Department of Health, Education, and Welfare, Public Health Services. Credentialing health manpower (DHEW Publication No. (05) 77-50057). Washington, D.C.: Author, July, 1977.

United States Equal Employment Opportunity Commission, Guidelines in employee selection procedures. Washington, D.C.: Aug. 24, 1966 (29 C.F.R.: 1607).

United States v. State of North Carolina Civil No. 4476 (E.D.N., CAR., 1975).

COMMON THEMES IN TEACHER CERTIFICATION TESTING PROGRAM DEVELOPMENT AND IMPLEMENTATION

Katherine E. Vorwerk

William Phillip Gorth

National Evaluation Systems, Inc.
30 Gatehouse Road
Amherst, MA 01004

A Paper Presented at the Annual Meeting
of the National Council on Measurement
in Education, New York, 1982

Overview

Programs of teacher competency training and teacher competency testing prior to certification are not new. Yet there still exists confusion about the intended purpose and outcomes of such programs, particularly teacher certification testing programs. For example, it has been said that teacher certification testing programs:

- will either improve the quality of education or lower the teaching profession's standards because of their emphasis on minimal knowledge;

- will either serve to define what a good teacher is or end up being nothing more than a "search for victims" and a "hollow means of judging the efficacy of teachers" (Cole, 1979); and

- will either test for content that is unrelated to successful teaching or test for content that is an absolute necessity.

As is true of all occupational licensing laws, the primary purpose of teacher certification laws and their testing component is to "protect the public health, safety, and welfare" by ensuring that only individuals who are competent in a subject are allowed to teach it. Yes, certification testing programs do in most cases emphasize minimum content knowledge; yes, they can result in improvement in the quality of education; yes, they may end up being

part of a definition of what a good teacher is and what content knowledge is absolutely necessary, etc. But these are secondary outcomes of such programs. The primary outcome, which every program is designed to achieve, is the protection of the public from incompetence.

The public is clearly concerned about teacher competence. For example, in a recent Gallup Poll, 95% of those polled agreed that teachers should be requested to pass exams in their subject areas (Cole, 1979). Teacher incompetence is frequently used by parents and legislators as a partial explanation for the decline in students' test scores that we have witnessed over the past 15 years. Moreover, the large number of states that require or soon will require a teacher certification testing program (approximately 15), or are considering doing so, is further testimony to the fact that the public wants its children protected from incompetent teachers.

A systematically developed teacher certification program can potentially prevent individuals who lack competence in critical subjects from entering the teaching profession. This paper presents a general model for developing the testing component of a certification program. The model's structure will be described and the key issues associated with each component of the model will be presented.

It should be pointed out that the model applies only to the formal testing component of a teacher certification program such as a structured observation session or a paper-and-pencil content test. It does not apply to other parts of a certification program such as course requirements or student teaching.

Development Model

The model consists of five components (see Figure 1) which are:

(1) Developing Certification Requirements;

(2) Deciding How to Assess Requirements;

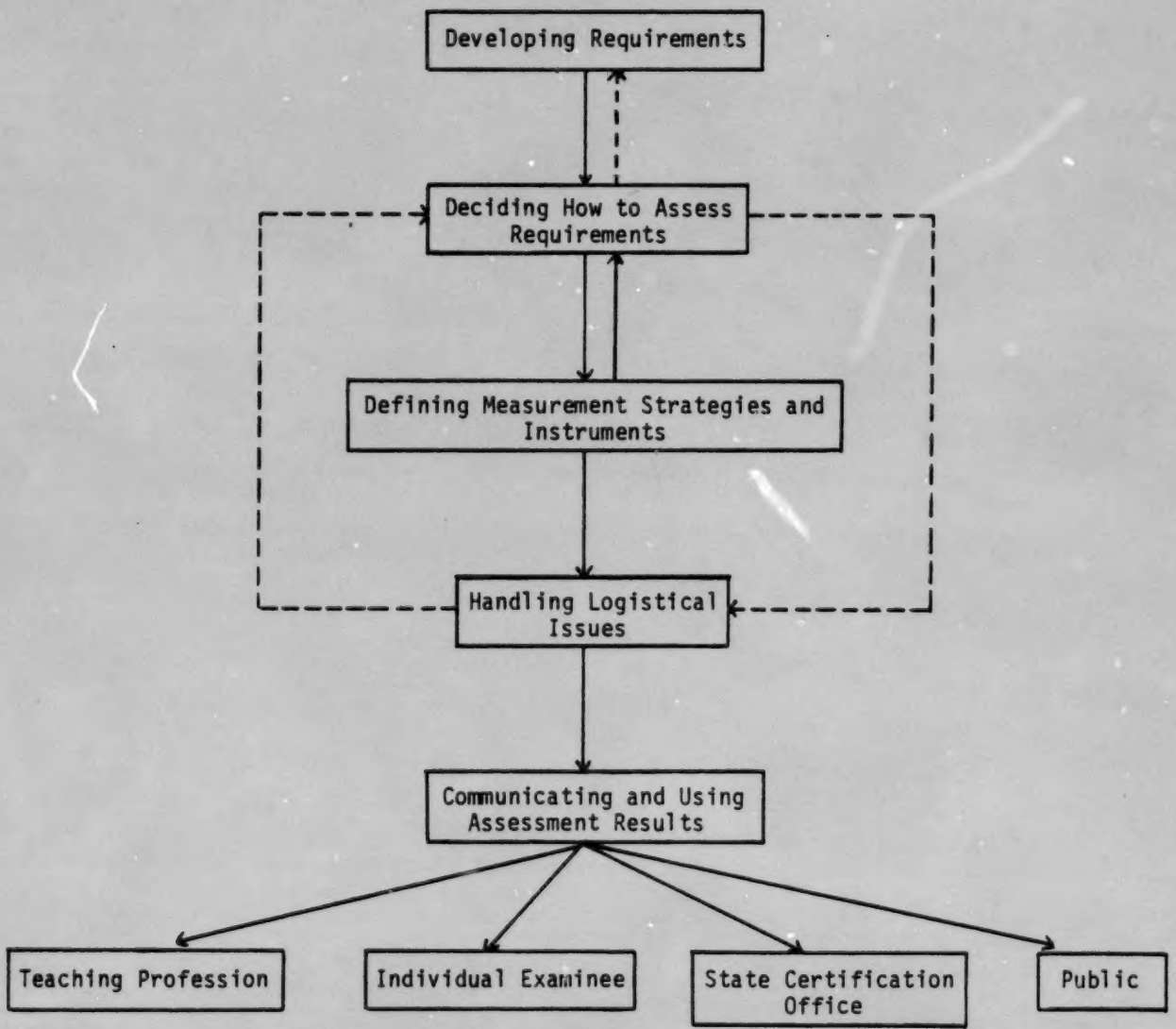
(3) Defining Measurement Strategies and Instruments;

(4) Handling Logistical Issues of Assessment; and

(5) Communicating and Using Assessment Results.

Figure 1

TEACHER CERTIFICATION TESTING
PROCESS MODEL



The components are roughly sequential, although several of the steps overlap. Each component will be discussed separately in the sections that follow.

Developing Certification Requirements. Requirements mandating teacher certification testing generally come either from state boards of education or state legislatures. For example, the authority for Alabama's testing program comes from the State Board of Education, while that for Florida comes from the legislature. Occasionally, the effort may be a joint one. After requirements are developed, they are generally passed on to the state's department of education for further definition and implementation. Ideally, the department of education provided input during the development of the mandated requirements, and therefore, is at least familiar with their content.

Other constituencies that should be represented in developing certification requirements are the state's teacher training institutions, teachers, and the general public. Each of these constituencies will be affected by the requirements. Their input during the initial stages of defining the requirements will help to ensure that the requirements are both workable and acceptable.

At this stage some states contact other states' departments of education for advice and background on their certification programs, or they engage the services of one or more testing consultants who also can provide information about existing certification testing programs as well as psychometric consultation.

Deciding How to Assess Requirements. Generally, the state's department of education is responsible for deciding how the requirements will be assessed. Requirements, of course, vary widely. Some merely specify the use of a particular test or series of tests; e.g., Arkansas's regulations "require persons applying for initial certification to satisfactorily complete 'an existing teachers' examination' or other similar examination." In such cases, the state department moves on to the next component of the model. In other states the requirements specify a more comprehensive and detailed testing program that could include entrance exams to teacher education programs, exit exams covering basic skills and specific content knowledge, and other nonexamination requirements (e.g., practice teaching and in-service training for certified teachers and administrators).

Deciding how to assess requirements impacts heavily on the measurement strategies that will be used as well as on the type of results that will be produced. For example, deciding to assess teaching skills using some type of on-the-job observation procedure will result in the implementation of a very different type of measurement instrument than if a decision is made to assess the content knowledge competency of teachers in the subject or subjects they aspire to teach.

In reaching a decision, several key issues must be considered. First, budgetary constraints must be realistically evaluated. Assessment strategies will vary in cost. The cost will be paid partly by the state for start-up costs and partly by the individual examinee for operating costs. Second, time considerations are critical. Often mandated requirements include an implementation date. The development of a certification testing program tailored to

the state's curriculum requirements involves more time than adopting an existing test. If a testing program must be produced within three to five months, this will have an impact on the type of assessment selected. Third and finally, the question of which assessment strategy will best protect the state's public must be asked. For example, should systematic observation of teachers on the job and paper-and-pencil tests of content knowledge both be used or is one or the other sufficient? Clearly, some combination of observation, paper-and-pencil tests, and preservice evaluation is preferable, but given time and budgetary constraints, is this possible? If not, which approach is actually going to meet the needs of the state in the most satisfactory manner?

Defining Measurement Strategies and Instruments. This component is closely tied to the previous one. Deciding how to assess requirements is, in effect, defining measurement strategies. However, the creation of assessment instruments involves additional technical work.

This component is a major one, and it generally consumes most of the start-up resources expended on the testing program. In this phase the actual testing instruments are developed. Professional and legal guidelines for tests used for certification purposes apply here and must be clearly understood and followed. Major issues covered by these guidelines include the need for job relatedness of the measurement instruments, test validity, test reliability, and the passing score or standard that is used. Specific technical considerations involved in each of these will be discussed in a later symposium paper (Nassif & Elliot).

In addition to adherence to professional and legal guidelines for test development and use, the department of education must take care to involve members of the teaching profession--both actual teachers and teacher educators--in the development of assessment instruments. The involvement of these individuals is critical to ensure the appropriateness of the test instrument to the state. Clearly, teaching professionals would be involved in a job analysis procedure carried out to establish the job relatedness of the content of a particular examination. For example, Oklahoma surveyed over 4,500 teachers as part of the process of establishing the job relatedness of Oklahoma's certification tests.

In addition, teachers and teacher educators also should be involved in all other steps in this component of the model, particularly if a customized paper-and-pencil test or other measurement instrument is being developed. For example, committees of teachers and teacher educators should be formed to review the domain of knowledge/skills to be included on the test, to review the results of the job analysis procedure (and to make judgments on how those results are to be used) and finally, to review the actual test items appearing on the test. Such reviews by teaching professionals are typical of the testing programs in states such as Georgia and Alabama.

Handling Logistical Issues of Assessment. Registering candidates for the assessment and actually carrying out the assessment, the two major parts of this component, are logistically demanding. Depending upon the program, this component can vary from two or three administrations of one test at three sites distributed across a state to a sequence of preservice and in-service evaluations coordinated with a test of content knowledge given several times a year.

It is important at this point to provide teacher certification applicants with complete information about the testing program requirements, administration procedures, and results reporting. This notification seems best acccomplished through a detailed registration bulletin which clearly specifies the state's certification laws and regulations and explains to the applicant his or her responsibilities and rights during the testing program. The bulletins should be widely distributed in the state through the teacher education institutions and the department's certification office. Other information also may be available. For example, if a test is developed from a set of content objectives, these content objectives should be made available to students, through libraries or teacher education programs, so that they can use them to prepare for the test.

Registration procedures should be as simple as possible to avoid mistakes and confusion. Information about registration procedures, deadlines, fees, and testing locations should be in the registration bulletin.

Regardless of how the registration materials are provided to examinees, it is important that they be provided in advance of the administration so that students have ample time to peruse them, send in registration forms, and still have time to change their registration if they choose.

Test administrations should be standardized and secure so that all applicants have the same opportunity to perform. Also, administrations should occur several times during a year so that applicants have ample opportunity to sit for the exams, and they should be spaced properly so that the results of one administration are reported to the candidate before the registration deadline for the next administration. In Oklahoma, for example, testing sessions are held four times per year, and students may take up to eight tests at each two-day session.

Communicating and Using Assessment Results. At a minimum, examination results should be reported to four constituents. First, results should be reported to individual examinees. Clearly, those who take the test should find out whether they passed. But in addition to information on whether they passed or failed the test, examinees also should be provided with diagnostic information; i.e., score reports should include information on how the student performed on each of the major content areas covered by the exam. This diagnostic profile of the student's strengths and weaknesses can serve as a springboard for additional growth. For example, a test taken in health and physical education might provide feedback to the student on how he or she performed on questions related to elementary physical education, physical development, and mental health.

Second, results should be reported to the colleges and universities at which the certification applicants received their educations. These results can provide institutions with two types of useful information: (a) how each of their students performed on the test, and (b) the performance of each of their individual teacher training programs. Content where students show consistent strength or weakness may indicate corresponding areas of strength and weakness in the training programs themselves. Such information can stimulate curriculum modification and the strengthening of the training programs.

Third, the state should receive results. Obviously, the state needs this information about individuals to determine whether certification should be granted or denied. Statewide data also provide information about how the total group of students has performed and allow for comparison among subgroups within the sample, for example, males vs. females.

Fourth and finally, results should be reported to the public. Results demonstrate to the public that only teachers found to be competent in those areas determined to be necessary have been certified, and that their children are being protected from incompetent applicants to the teaching profession.

That is the model. As indicated at the beginning of this paper, there is time only to cover the model in a very general way. It has not been possible to discuss many of the important details involved in the definition and implementation of a teacher certification testing program. However, it has been possible to at least mention most of the important issues that need to be considered.

Conclusion

Teacher certification testing programs do not solve all of the problems of American education, or even the smaller cluster of problems related specifically to teacher competency. However, these programs are able to identify people who can and cannot demonstrate, in a relevant testing situation, the competencies which the state feels they should be able to demonstrate. The model presented in this paper provides a general description of the steps in the development and implementation of such a teacher certification testing program, and the issues that must be considered.

References

Cole, R. W. Minimum-competency tests for teachers: Confusion compounded. Phi Delta Kappan, December 1979, 61(4), 233.

Haefele, Donald L. How to evaluate thee, teacher - let me count the ways. Phi Delta Kappan, January 1980, 349-352.

Leiser, B. M. Incompetent teachers and misguided courts. National Forum, Spring 1981, 47-48.

Piper, M. K., & Houston, R. W. The search for teacher competence: CBTE and MCT. Journal of Teacher Education, September-October 1980, 31(5), 37-40.

Vlaanderen, R. B. Trends in competency-based teacher certification. Unpublished manuscript, Education Commission of the States, Denver, Colorado, 1980.

Vlaanderen, R. B. Testing for teacher certification: Summary chart. Unpublished manuscript, Education Commission of the States, Denver, Colorado, 1981.

VARIATIONS IN APPROACHES TO ASSESSMENT

Michael Priestley

National Evaluation Systems, Inc.
30 Gatehouse Road
Amherst, MA 01004

A Paper Presented at the Annual Meeting
of The National Council on Measurement
in Education, New York, 1982

Introduction

Since the mid-70s, concerns about the quality of teaching in public schools have led to significant changes in the requirements for teacher certification in programs throughout the country. An increasing number of states and school districts are looking at various methods of assessment which may help to improve the efficacy of the certification process (Harris, 1981; Nothern, 1980). Since 1978, five states have substantially renovated their programs to incorporate competency-based, criterion-referenced tests and performance assessment for evaluating teachers seeking initial certification (Note 1). Dozens of other states have begun the process of exploring options and implementing similar changes; still others require candidates for initial certification to pass some component(s) of the National Teacher Examinations (NTE).

A cry no less vocal than the call for teacher testing is the protest that no examination can adequately measure the skills essential to competent teaching (NEA, 1982). This perspective seems to posit that most or all teacher competencies are intangibles--words that begin with capital letters such as Patience and Enthusiasm. While it seems fairly apparent that no test of multiple-choice questions can suffice as the sole criterion for certification, it is also apparent that some form of content-based assessment is essential to ensure that candidates at least know the information they are supposed to impart in the classroom. Whether or not they can impart it successfully is

the subject of later assessment through different procedures. The American Federation of Teachers, for one example, supports the use of tests to assess the qualifications of candidates for certification, but not for decisions related to retention, salary, and tenure (Note 2).

Thus, the goal should not be to eliminate assessment and leave teacher training institutes on their own to maintain standards but to support the effort by improving the tests and other assessment methods available to evaluate teacher candidates for certification. This perspective naturally raises some significant conceptual issues which must be carefully considered.

Conceptual Issues

The first major issue to consider is when to assess teacher candidates. Recently developed programs seem to indicate agreement that prospective teachers should be assessed at at least two of three different stages (Note 3): qualifications for admission to a teacher training program, qualifications achieved upon completion of the program, and performance in the classroom. A comprehensive program for initial certification would provide assessment of teacher candidates at all three of these stages.

The second major issue is how to assess teacher candidates at each stage. On this issue alternatives abound, but agreement founders.

Assuming that assessment occurs at the three points mentioned, the third major issue to consider is how to conduct the assessment in a manner that is technically and legally defensible. According to federal employment guidelines, which also affect certification procedures, any instrument used for licensing or selection must be a representative measure of the actual domain

of skills used on the job. It must also be able to be validated for its actual or intended purpose (EEOC, 1978). In addition, state and local laws which apply specifically to certain programs or aspects of teacher certification must be heeded judiciously. In many cases state legislation or a board of education has provided the impetus for developing and implementing a teacher certification program. For example, state laws requiring competency tests have been passed in Florida, Oklahoma, South Carolina, and Texas; board of education mandates have been established in Alabama, Georgia, and New York.

The purpose of this paper is to explore various approaches to assessment for initial teacher certification. Conceptual issues and the relative merits of each approach currently available are considered in relation to test design, assessment for entry to a teacher education program, exit credentialing, and classroom performance assessment.

Assessment Design

Certification Areas

To a large extent, the first step in designing assessment instruments depends upon the structure of the state's certification program, i.e., the definition of certification areas. While one state may certify a teacher only in a general area called Social Studies, for example, another may certify teachers according to specialty: History, Political Science, Economics, and so on. The definition of these areas will influence the number and type of assessment instruments required. The first state would require only one general content-based test for the Social Studies certificate; the second

would have to develop an umbrella test for Social Studies and/or a discrete test for each of 6-8 specialty areas. The major reason for this is to ensure that each candidate is only responsible for content essential to his or her field; e.g., a person who would be certified only to teach Economics should not be required to pass a test that includes U.S. History and Geography.

The definition of tests measuring a specific array of certification areas usually precedes, but often depends on, determining what to measure within each test. One important fact to keep in mind: tests should be developed or adapted to certification areas--not the other way around--in order to maintain the integrity of the state's own program design.

Domain Definition

Determining what to test for admission, for initial certification, and for classroom performance assessment involves defining domains of knowledge and skills for each assessment area. Assessing qualifications for admission to a teacher education program may involve an evaluation of the student's academic records or a test of basic skills, literacy, and communication. Exit requirements may involve another evaluation of the candidate's credentials, a test of content knowledge in a chosen teaching field, a test of pedagogy, or alternative assessments of various performance skills. Evaluating performance in the classroom may involve any of a large number of assessment strategies.

In the process of designing a comprehensive assessment program, the task of determining what to assess must precede or occur at the same time as choosing assessment methods. Basically, there are two ways of determining what to assess.

One method is to identify the knowledge and skills taught at the college level. For example, a candidate for teacher certification could be tested on his or her knowledge of the curriculum required by the teacher education program. This can be a legally defensible method (Note 4), according to the notions of "curriculum" and "instructional" validity, and it seems fair to the candidates: they are tested only on what they have been taught in teacher training. However, it may not be fair to students in the classroom because this approach assumes that colleges instruct teachers in what they need to know in order to teach. What teachers actually have to know in order to teach in the classroom may differ from what the colleges have prepared them to teach.

A second method--job analysis--solves this problem and lays the foundation for establishing that the test measures a representative sample of knowledge and skills required on the job, in accord with federal guidelines. In teacher certification programs, job analysis has been used successfully in several states, including Georgia, Alabama, and Oklahoma.

Essentially, a job analysis--conducted by survey, observation, and/or interviews--generates empirical data describing what people do in their jobs, thereby identifying the qualifications needed of a candidate who wants to be certified for that kind of job. In one approach to job analysis for teachers, skills and content knowledge are defined by behavioral objectives, which are rated by job incumbents (practicing teachers) as to their job relatedness (time spent teaching or utilizing the content of the objective and its essentiality).

From the results of the ratings, the objectives can be rank ordered by these dimensions across the overall list and can be ordered within "subareas" used to group the objectives. When selecting objectives for assessment, it is important to select the <u>most</u> job-related objectives in each subarea. This

ensures that the selected objectives reflect the proportional size of each subarea in relation to the size of the total job-related field. In turn, this proportionality design provides an initial estimate of a blueprint or structure for the assessment instrument(s), which can be developed to reflect the relative importance of each subarea containing job-related objectives.

Using a job analysis to define assessment domains provides an empirical basis for developing the instruments. However, a certification program in a given state should meet additional concerns: as the NEA (1982) points out, teachers must and should have considerable involvement in the assessment process. Among other roles, constituency groups can help to identify emerging fields, which teachers may not teach now but may have to next year or the year after (e.g., metrics, the use of calculators); they can ensure that the assessment instruments serve the intended focus of education within the state; and they can ensure that the language and structure of the content is appropriate to the region (e.g., one state might teach the theory of evolution, while another requires a different approach).

Empirical information from a job analysis and expert judgments from teachers and other constituencies can provide the foundation for determining what to assess. The next step is to determine how to assess the competencies identified as essential for entry, initial certification, and classroom performance.

Assessment Methods

Entry Tests

Assessment of qualifications for admission to a teacher education program usually occurs in the candidate's second year of college study, prior to entry into the program at the beginning of the third year. Traditional methods of assessing such qualifications have most often included teacher recommendations of the student candidate and an examination of the student's academic record (grades, course requirements, etc.). However, recently developed programs in South Carolina and Alabama have abolished this essentially pro forma approach; instead, they require statewide entry tests to ensure that candidates have the basic skills (e.g., mathematics, communication skills, general education) required for some degree of success in the teacher education program. Another possibility, recommended by Watts (1980), is to establish a professional standards board for admissions that functions independently of training institutions (as some other professions have done, e.g., engineering, architecture).

Of these three approaches, the most efficient means of assessing entry qualifications appears to be some form of entry test. Whether the qualifications are identified as general education (i.e., liberal arts) or literacy and basic skills, the entry test may involve assessment methods other than strictly multiple-choice, paper-and-pencil tests. The key is to decide what to test, then how to test it most effectively and efficiently. If entry qualifications include literacy skills of reading, writing, and listening, for example, then assessment methods must be capable of measuring the skills required.

South Carolina has begun developing a basic skills entry test of reading, writing, and mathematics, the content of which was identified through an extensive survey. Alabama has already developed and implemented an entry exam called the English Language Proficiency Test, which was administered for the first time in November, 1981. Its content, derived from a validation survey of practicing teachers in all fields, includes reading, writing, language skills, and listening. Methods of assessing these areas are listed in the chart that follows:

Alabama's English Language Proficiency Test

Content Area		Assessment Method
Reading	--	A "cloze" test of reading comprehension, using multiple-choice items with 5 choices
Writing	--	An essay test, scored by the holistic method
Language Skills	--	A multiple-choice test (4 choices per item) of basic grammar, mechanics, and reference skills
Listening	--	A listening tape of passages read aloud, testing comprehension by multiple-choice items

In addition to these two state-developed programs, nationwide standardized tests are also available. One possibility is the "Common" examination portion of the NTE, which currently includes a section on Professional Education and General Education (social studies, literature and fine arts, science, etc.). According to a recent announcement by Educational Testing Service (Note 5), the "Common" portion of the exam will be revamped. The new version will essentially include all of the present components, plus a new section on Communication Skills (listening, reading, writing).

Exit Tests

The next stage in the process is the assessment of qualifications for initially certifying a teacher who has completed a teacher training program. Once the qualifications have been specified, by job analysis or other method, a number of options exist for assessing these qualifications. Past certification procedures have been based largely on the candidate's completion of an accredited teacher preparation program (Hathaway, 1980). But this approach necessarily assumes adequate standards of competency enforced by each program and some relative comparability across programs in a given state. With the increasing concerns for the actual competency level of teachers, fewer states are willing to assume the quality of teacher preparation programs; more and more states have implemented, or will implement, teacher competency tests for initial certification.

In most areas the essential competencies are content based, thus measurable by paper-and-pencil tests. For these areas, states have the option of selecting and adopting an existing standardized test if it meets their needs; developing their own state-specific tests; or achieving some combination of the two approaches.

Basic or professional skills tests. Exit tests are often used to assess basic skills or professional skills which are common to all teaching areas. For example, Tennessee has used the California Achievement Test as an exit test of basic skills for all teacher candidates; many states have used one or more portions of the NTE's Common Examination to measure general education and/or professional skills. In some states (e.g., Washington), colleges and universities develop and require their own tests of general education.

Three states, Florida, Alabama, and Arizona, have developed their own instruments. Florida's Teacher Certification Examination, based on a list of 23 generic competencies, measures reading, writing, mathematics, and professional education. Assessment methods include a multiple-choice cloze test, an essay test scored holistically, and tests of multiple-choice items. Alabama requires that all teacher certification candidates pass a multiple-choice test of basic professional studies, which is based on a job analysis of practicing teachers in all fields. In addition to this test, candidates in Alabama must pass content knowledge tests specific to their teaching areas. In Arizona's program, currently under development, all teachers will have to pass a multiple-choice test of generic teaching knowledge and skills.

One important consideration in choosing assessment methods for professional education or pedagogy tests required of teacher candidates in all fields is to distinguish between content knowledge, which is measurable by paper-and-pencil tests, and classroom skills, which are not directly measurable by this technique. Professional skills required in the classroom should be measured during the performance assessment stage of the certification process.

Teaching field tests. Exit tests are also used to measure teaching field content knowledge. Here again, states have several options for selecting or developing tests for this purpose. Choosing the NTE, which provides tests of some 26 specific areas, has its advantages and disadvantages. On the positive side, the NTE is relatively inexpensive (compared to the cost of developing new tests). Also, it can be adopted and implemented in a relatively short time--an important factor if state law or mandate requires rapid implementation. On the negative side, adoption of the NTE can pose some potential problems. First, it has a preestablished set of teaching area tests available; thus, a state

must adapt its certification areas to the test and adopt some other method of certifying in areas not covered. Second, to conform to legal requirements, the NTE must usually be validated within the state where it will be used (a process which can take up to three years). That is, the content of the tests must be compared and analyzed empirically in relation to state teacher preparation curriculum (as has occurred in South Carolina and will occur in Virginia) or to the results of a teacher job analysis. Thus, adoption is not as straightforward and unencumbered as it may seem. Third, the NTE provides normative-referenced scores comparing a student's performance to the performance of others; or, in some modified programs, to standard scores determined within a state. The student receives a pass or fail and a numerical score but (unless special modifications are made) no indication of strengths and weaknesses, which could be extremely helpful both to the institutions and to the students who must retake the test(s).

If adoption of available tests does not satisfy a program's requirements, then a second alternative is to develop new tests. In the past four years, several states have developed their own content area tests for initial teacher certification to meet their own specific needs. Georgia began in 1975 and has since implemented tests in 23 different areas, all of them based on extensive job analysis and teacher involvement. Tests in eight more fields are currently under development. Alabama has developed tests in 31 areas, which were administered for the first time in December, 1981. Oklahoma has developed tests covering 79 different areas: 26 general tests for individual teaching fields; 8 umbrella tests for such fields as Social Studies, Mathematics, and Language Arts; and 45 specific area tests, which must be taken along with the appropriate umbrella exam(s).

The advantages of full-scale developmental efforts are manifold: the criterion-referenced tests are based on job analyses; the tests match the state's certification areas; they can be empirically content validated; and test scores provide indications of relative strengths and weaknesses on specific domains within each test. In addition, teachers and administrators within the state participate in the development process, thereby ensuring the relevance of the tests and helping to instill grassroots support for the testing program. As to disadvantages, the first is cost: large-scale development projects can be expensive. The second, which only applies in some cases, is the time required for development. If done properly, programs of such magnitude and complexity require anywhere from one and a half to four years for development, which may be a disadvantage if a mandate has limited the time available.

But what if neither of these alternatives--adoption or development--is suitable? Some states have combined certain aspects of each alternative to create specially tailored certification programs. For example, South Carolina is currently in the process of developing teaching area tests in ten specific fields. Other certification areas offered in South Carolina require the candidate to take specified portions of the NTE. However, since a state law requires the reporting of strengths and weaknesses for all teacher tests, the NTE's normal scoring method must be altered to suit South Carolina's requirements.

Special concerns. While these procedures for selecting or developing paper-and-pencil tests may seem relatively straightforward, a number of special concerns will arise during the process. The first, mentioned earlier, is the design of certification areas: general tests, special area tests, and so on.

In a field such as special education, this can be a critical and volatile issue. The second, also related to test design, is the need to provide subtest or domain scores. This provision requires careful design to ensure adequate measurement of knowledge and skills not just within the test as a whole, but also within each subtest or domain. The third special concern, related to assessment methods, deserves more detailed exploration at this point: multiple-choice, paper-and-pencil tests may not adequately cover the representative domain of content knowledge in some teaching fields. While some special fields may be "low incidence" (i.e., only a few people certified annually), thus delegated to local assessment programs, they must be considered first at the state level. Several examples here may be helpful.

The content knowledge required of a prospective teacher in music or a foreign language may only be partially covered by a paper-and-pencil test. Music teachers must also be able to listen to and recognize musical selections, proper articulation, misplayed notes, and so on. For this reason, both Georgia and Oklahoma have developed listening tests in music: examinees listen to the tapes, then answer multiple-choice questions. Similarly, a foreign language teacher must be able to speak and to understand the language; thus, tests such as the NTE and those developed in Alabama, Georgia, Oklahoma, and South Carolina include language-tape tests. However, in most cases, speaking tests occur at the local rather than state level.

In vocational areas, on-the-job performance is often essential to the teacher candidate's preparation. The area commonly called "Trades and Industries (T & I)," for example, includes trades as diverse as cosmetician, tailor, and diesel mechanic. Most states require a T & I teacher to be licensed and experienced within the trade he or she would teach; they also may

require some amount of teacher training. Tests in the T & I field, as in South Carolina, for example, may include both a paper-and-pencil test of the generic skills taught in teacher training to all T & I candidates and an actual or simulated performance test of trade skills (conducted by the colleges themselves).

In summary, these kinds of special concerns will undoubtedly arise in the development of a certification program. Efforts to accommodate these concerns must consider the use of alternative assessment methods to measure skills which cannot be assessed adequately by strictly multiple-choice, paper-and-pencil tests; at the same time, they must consider the cost and practicality of alternative methods (Priestley, 1982).

Classroom Performance Assessment

The third and final stage of initial teacher certification is the assessment of classroom performance, usually conducted during the period in which the candidate holds a temporary or "provisional" license or certificate. The basic goal of performance assessment is twofold: (1) to help the teacher improve his or her skills, and (2) to collect information on which to base an administrative decision as to whether or not the candidate should receive full or "permanent" certification. Scriven (1981) distinguishes between the assessment methods appropriate to these goals by identifying the requirements and benefits of formative and summative evaluation.

Achieving these goals demands that assessment of performance be limited to competencies that teachers would be expected to possess as entry-level professionals, and that the assessment methods provide fair, reliable measures of

competencies determined to be essential. The first demand is a function of defining the domain of essential competencies, a process that may be based on teacher training curriculum or on job analysis (as stated earlier in relation to content-based tests). The second demand, for adequate assessment methods, requires a broader perspective.

As MacDonald (1973) reported, the state of the art of performance assessment technology was a "rather depressing picture" in 1973. Since then, however, considerable progress has been made as the demand for more effective methods has become more clamorous and persistent. Unlike at the first two stages--entry and exit tests--assessment at this third stage does not include the option of standardized, off-the-shelf instruments for performance assessment. On the other hand, the methods available are numerous, and there are programs to consider as potential models for the development of performance assessment procedures. Most important at this stage is the development of an assessment that meets the specific needs of a state or local program, at the level on which actual evaluations will occur.

In terms of methods for assessing teacher performance, Medley (1978) constructively proposes six general alternatives, and Haefele (1980) critically reviews twelve (with considerable overlap among the alternatives presented). Millman (1981) examines a number of methods in depth, with relation to their use in teacher evaluation, and many of these methods can be adapted for use in assessment for initial certification.

Simply classified, the methods of assessment involve three basic types: observational ratings of the teacher in action (by students, peers, supervisors, principals, independent evaluators); training/simulation exercises; and

testing of the teacher's classroom students (e.g., before and after instruction). Within each of these categories are a number of specific techniques, but some are more useful than others in measuring performance on specified competencies. For example, as Medley (1978) points out, Popham's (1975) suggested approach of the "teaching test" and related approaches involving pre- and post-tests of the classroom students can really only yield overall means and test scores. While these kinds of data might be useful, they cannot be matched directly to specified teacher competencies. Assessment of particular skills identified as essential to adequate teacher performance requires the use of methods that can measure and provide feedback on each skill, for both formative and summative needs.

Programs designed to accomplish these purposes have been developed in South Carolina and Georgia. In the Georgia program, in addition to meeting the requirements of course credits and grades, and passing a criterion-referenced teaching area test, the teacher candidate undergoes performance assessment during the first year while holding a provisional certificate.

Georgia's Teacher Performance Assessment Instruments (Note 6) were developed to measure performance in relation to specific teaching skills identified through an extensive survey as both generic and essential to teaching in all fields. Assessment is governed and provided by five different instruments, as described below:

Instrument	Method of Assessment
• Teaching Plans and Materials Instrument (TPM)	-- A portfolio of instructional preparation rated by data collectors who also interview the teacher
• Classroom Procedures Instrument (CP)	-- Direct classroom observation of teaching methods and practices
• Interpersonal Skills Instrument (IS)	-- Direct classroom observation of the teacher's ability to create a sociable atmosphere and manage classroom interactions
• Professional Standards Instrument (PS)	-- Interviews with the teacher, his or her colleagues, and supervisor to gather information on professional conduct (complying with policies and procedures, participating in professional growth activities, etc.)
• Student Perceptions Instrument (SP)	-- A questionnaire filled out by students, composed of items parallel to those in the CP and IS instruments

For each of the first four instruments, at least three trained data collectors (peers, supervisors, principals, independent evaluators, et al.) rate the teacher's performance on each indicator on the basis of a 5-point scale. Mean scores across all raters and all indicators are calculated by computer or by hand.

It is important to note here that only the first three instruments are used for summative certification decisions; the other two--student perceptions and professional standards instruments--are used formatively to determine the need for in-service training and to create teacher performance profiles.

Conclusion

This paper, in relation to several conceptual issues, has explored a number of assessment options for initial teacher certification. A basic tenet stated at the outset is that assessment should occur at three stages: before admission to a teacher training program, upon completion of the program, and during on-the-job performance in the classroom. Certification should be based on at least these three assessments and not on any one of them as the sole criterion.

Regarding the assessments themselves, the content or domain of what to assess should be defined carefully, preferably through job analysis and with extensive teacher involvement. Assessment instruments should then be designed and either selected or developed to measure the specified domains as effectively and efficiently as possible. Above all, given the recognition and acknowledgment of the fact that state needs, teacher training programs, and qualifications for different teaching fields vary considerably, all assessment methods should be fitted to the specific needs of a given situation. No one all-encompassing solution is possible for assessing competence in a profession of such importance, variation, and frequent change.

Reference Notes

1. The five states are Alabama, Florida, Georgia, Oklahoma, and South Carolina.

2. American Federation of Teachers, AFL-CIO Convention Resolutions, 1979. Washington, D.C.: American Federation of Teachers.

3. Programs developed in Alabama, Georgia, Oklahoma, and South Carolina all incorporate two or three of these components.

4. United States v. State of South Carolina Education Division, U.S. District Court, Civil Action No. 75-1610, April, 1977.

5. Unpublished program description disseminated by Fred Harris of Educational Testing Service, Princeton, New Jersey, 1981.

6. Johnson, C., Ellett, C., and Capie, W. An Introduction to the Teacher Performance Assessment Instrument: Their Uses and Limitations. Athens, Georgia: Teacher Assessment Project, College of Education, University of Georgia, 1980.

References

Haefele, D.L. How to evaluate thee, teacher--Let me count the ways. Phi Delta Kappan, January 1980, pp. 349-352.

Harris, W.W. Teacher command of subject matter. In Millman, J. (Ed.) Handbook of Teacher Evaluation. Beverly Hills: Sage Publications, 1981.

Hathaway, W.W. Testing teachers to ensure competency: The state of the art. A paper presented at the annual AERA convention, Boston, April 1980.

MacDonald, F.J. The state of the art in performance assessment of teaching competence. A paper presented at the annual AERA convention, New Orleans, 1973.

Medley, D.M. Alternative assessment strategies. Journal of Teacher Education, March-April 1978, Vol. XXIX, No.2, pp.38-42.

Millman, J. (Ed.) Handbook of Teacher Evaluation. Beverly Hills: Sage Publications, 1981.

National Education Association. A closer look at teacher competency testing. NEA Reporter, January-February 1982.

Nothern, E.G. The trend toward competency testing of teachers. Phi Delta Kappan, January 1980, p. 359.

Popham, W.J. Performance tests of teaching proficiency: Rationale, development, and validation. American Educational Research Journal, 1975.

Priestley, M. Performance Assessment in Education and Training. Englewood Cliffs, NJ: Educational Technology Publications, Inc., 1982.

Scriven, M. Summative teacher evaluation. In Millman, J. (Ed.), Handbook of Teacher Evaluation. Beverly Hill: Sage Publications, 1981.

Watts, D. Admission standards for teacher preparatory programs: Time for a change. Phi Delta Kappan, October 1980, pp. 120-122.

TEACHER CERTIFICATION TESTING

TECHNICAL CHALLENGES: PART I

Paula M. Nassif

National Evaluation Systems, Inc.
30 Gatehouse Road
Amherst, MA 01004

A Paper Presented at the Annual Meeting
of The National Council on Measurement
in Education, New York, 1982

Introduction

Teacher certification testing programs present challenges to the practitioner regarding several technical issues. Parts I and II of this paper will focus on standard setting and equating, and validity and job analysis, respectively. A review of these technical issues requires a delineation of the methods currently in use. In each section that follows, present approaches will be described and discussed. Recommendations for alternatives will be suggested where appropriate.

Standard Setting

Clearly one of the most significant aspects of tests developed and used for employment decisions is setting the passing score or cut score. This area of research is a broad field of its own--replete with legal factors, technical concerns, and logistical considerations. There are several models available for standard setting. Koffler (1980) and Hambleton & Eignor (1978), among others, have studied various methods and examined their appropriateness, accuracy, and usefulness. Many methods of standard setting have been used frequently in student competency assessment. These methods include: Nedelsky (1954), Angoff (1971), Ebel (1972), Jaeger (1978), Contrasting Groups and Borderline Groups (Zieky & Livingston, 1977).

However, when one reviews the methods actually used in setting cut scores for teacher certification testing, one finds a smaller list than the one above.

In the past 15 years, state-mandated use of the National Teacher Examinations (NTE) often involved the administration of the exam and the establishment of a passing score by state administrative decision. The procedure was not empirical, nor did it result in a cut score that systematically bore relationship to successful performance on the job. In some states the use of the NTE with an arbitrarily set cut score was legally challenged. In the following cases the continued use of the exam in this manner was not allowed by law: United States v. North Carolina (1975), Baker v. Columbus Municipal Separate School District (1976), and Georgia Association of Educators v. Jack P. Nix (1976). In a case that involves the use of a cutoff score to determine those candidates that are qualified or unqualified, the user of the test must give sufficient proof that the cutoff was not established in a capricious or arbitrary manner.

In South Carolina in 1977, it was found that the use of the NTE resulted in adverse impact against blacks. However, the state decided to investigate the test, validate it in South Carolina, and set cut scores in a systematic, empirical fashion. The result was that some of the NTE tests were validated and approved for use in South Carolina. This situation in South Carolina is a blend of using an "off-the-shelf" test and a test customized for state use. This is discussed further.

Approaches

Underlying most methods used are the procedures designed by Nedelsky (1954) and Angoff (1971). These procedures have been modified, consolidated, lengthened, and abbreviated for use in several states. They are described and discussed below.

Nedelsky (1954). Nedelsky (1954) has outlined the procedure as it would be used by instructors reviewing multiple-choice items to set a standard for a classroom test.

Description of the Technique

Letter grades F, D, C, B, and A used in this article have the conventional meaning of failure (F), barely passing (D), etc.

The proposed technique for arriving at the minimum passing score of an objective test, each item of which has a single correct response, is as follows:

Directions to Instructors

Before the test is given, the instructors in the course are given copies of the test, and the following directions:

> In each item of the test, cross out those responses which the lowest D-student should be able to reject as incorrect. To the left of the item write the reciprocal of the number of the remaining responses. Thus if you cross out one out of five responses, write 1/4.
> Example. (The example should preferably be one of the items of the test in question.)
>
> Light has wave characteristics. Which of the following is the best experimental evidence for this statement?
>
> A. Light can be reflected by a mirror.
> B. Light forms dark and light bands on passing through a small opening.
> C. A beam of white light can be broken into its component colors by a prism.
> 1/4 D. Light carries energy.
> E. Light operates a photoelectric cell.

Preliminary Agreement on Standards

After the instructors have marked some five or six items following the directions above, it is recommended that they hold a brief conference to compare and discuss the standards they have used. It may also well be that at this time they agree on a tentative value of constant k (see section on The Minimum Passing Score). After such a conference, the instructors should proceed independently.

Terminology

In describing the method of computing the score corresponding to the lowest D, the following terminology is convenient:

a. Responses which the lowest D-student should be able to reject as incorrect, and which therefore should be primarily attractive to F-students, are called F-responses. In the example above, response E was the only F-response in the opinion of the instructor who marked the item.

b. Students who possess just enough knowledge to reject F-responses and must choose among the remaining responses at random are called F-D students, to suggest borderline knowledge between F and D.

c. The most probable mean score of the F-D students on a test is called the F-D guess score and is denoted by M_{FD}. As will be shown later, M_{FD} is equal to the sum of the reciprocals of the numbers of responses other than F-responses. (In the example above, the reciprocal is 1/4.)

d. The most probable value of the standard deviation corresponding to M_{FD} is denoted by σ_{fd}.

It should be clear that "F-D students" is a statistical abstraction. The student who can reject the F-responses for every item of a test and yet will choose at random among the rest of the responses probably does not exist; rather, scores equal to M_{FD} will be obtained by students whose patterns of responses vary widely.

The Minimum Passing Score

The score corresponding to the lowest D is set equal to $\bar{M}_{FD} + k\sigma_{FD}$, where $\bar{M}_{FD}$ is the mean of the M_{FD} obtained by various instructors, and k is a constant whose value is determined by several considerations. The F-D students are characterized not so much by the positive knowledge they possess as by being able to avoid certain misjudgments. Most instructors who have used the F-D guess score technique have felt that this "absence of ignorance" standard is a mild one, and that therefore the minimum passing score should be such as to fail the majority of F-D students. Assigning to k values -1, 0 ,1, and 2 will (on the average) fail respectively 16 percent, 50 percent, 84 percent, and 98 percent of the F-D students. An informed final decision on the value of k can be reached after the instructors have chosen the F-responses, for at that time they are in a better position to estimate the rigor of the standards they have been using. In keeping within the spirit of absolute standards, however, the value of k should be agreed on before the values of M_{FD} are computed and certainly before the students' scores are shown.

It is the essence of the proposed technique that the standard of achievement is arrived at by a detailed consideration of individual items of the test. Only minor adjustments should be effected by varying the value of k. The reason for introducing constant k, with the attendant flexibility and ambiguity, is that F-responses in most examinations vary between two extremes; the very wrong, the choice of which indicates gross ignorance, and the moderately wrong, the rejection of which indicates passing knowledge. If a particular test has predominantly the first kind of F-responses, this peculiarity of the test can be corrected for by giving k a high value. Similarly, a low value of k will correct for the predominance of the second kind of F-responses. It is expected that in the

majority of cases a change of not more than $\pm.5$ in the tentative value of k agreed upon during the preliminary conference should introduce the necessary correction. It would be difficult to find a theoretical justification for values of k as high as two; for most tests the value of $k = 0$ is probably too low. This suggests a rather narrow working range of values, say between .5 and 1.5 with the value $k = 1$ as a good starting point.

If a part A of a given test consists of N_A items, each of which has s_A non F-responses (one of these being the right response), the F-D guess score for each item; i.e., the probability that an F-D student will get the right answer in any one item, is $p_A = 1/s_A$. The most probable values of the mean and the square of the standard deviation on this part of the test are given by $M_A = p_A N_A$ and $\sigma_A^2 = p_A (1 - p_A) N_A$. M_{FD} and σ_{FD} for the whole test are given by $M_{FD} = \Sigma_A M_A$ and $\sigma^2_{FD} = \Sigma_A \sigma^2_A$. The value of M_{FD} must be accurately computed for each test. σ_{FD}, however, may be given an approximate value. In a test of five-response items s may vary from one to five. If these five values are equally frequent, $\sigma_{FD} = .41\sqrt{N}$. If, on the other hand, the extreme values, $s = 1$ and $s = 5$, are less frequent than the other three values, as seems likely to be true for most tests, $.41\sqrt{N} < \sigma_{FD} < .50\sqrt{N}$. Since $K\sigma_{FD}$ is usually much smaller than M_{FD}, approximations are in order. With $k = 1$ and $\sigma_{FD} = .45\sqrt{N}$, the equation, Minimum Passing Score = $M_{FD} + .45\sqrt{N}$, should work out fairly well in the majority of cases and is therefore recommended as a starting point in experimenting with the proposed technique.

Refinements of the Technique

The definition of the F-response given above has an element of ambiguity. The lowest D-student may be expected to reject a given response on its own merits as clearly incorrect or because it is

clearly less correct than some of the other responses. In the example given under "Directions to Instructors" response E cites evidence against the wave theory of light and thus is an F-response on its own merits; other responses are consistent with the theory and may be considered non F-responses. It may be argued, however, that even a D-student should see that response D constitutes less cogent evidence than some of the other responses, and that therefore it is an F-response. Judging a response in comparison with other responses is theoretically sound, for it probably more closely corresponds to the mental processes of the student. To make a proper judgment of this kind requires time and considerable pedagogical and test-wise sophistication; with responses more heterogeneous than in the example cited a reliable judgment may be impossible. Experimentation with both definitions of the F-response is certainly in order, but at least in the beginning, the simpler version, i.e., judging each response on its own merit, is to be preferred.

Some instructors find it difficult in a good number of cases to decide whehter a response is an F-response. There is no theoretical reason against assigning to such a response half the statistical value of an F-response. (If, in the example cited, response D has been assigned the value of 1/2, the item would have had 1.5 F-responses and 3.5 non F-responses. Consequently the value of p for the item would have been 1/3.5 rather than 1/4.) If methodically and conscientiously pursued, such a procedure may result in a better agreement among the instructor's. It is not recommended as a substitute for clear and hard thinking about the degree of correctness of a response.

In theory, the proposed technique can be extended to assigning minimum scores corresponding to grades C, B, and A. The author has few data bearing on such an extension; they indicate fairly clearly, however, that a very thorough discussion of the meaning of the grades

of C, B, and A among the participating instructors must precede actual marking of the test. It seems fairly certain, moreover, that even if the instructors reach a really circumstantial verbal agreement on the meaning of these grades, modifications of the proposed technique are likely to be necessary. For, though an "absence of ignorance" standard may be adequate for identifying the barely passing students, more positive indications of achievement corresponding to higher grades seem desirable.

Perhaps a reasonable D-C guess score can be obtained by requiring the lowest C-students to reject responses that are in certain respects or, to a certain degree, inferior to other responses; the kind and the degree of inferiority must, of course, correspond to the instructors' definition of the meaning of the grade of C. To establish minimum scores corresponding to grades B or A, an instructor should probably focus his attention on the correct response and inspect the wrong responses primarily for their degree of deviation from the correct response; the allowable deviations for the lowest B or A will depend on the meanings assigned to these grades.

As the preceding paragraph suggests, the criteria used for determining the minimum scores corresponding to lowest D, C, and B or A may be qualitatively different; the method for computing these scores may be the same for all grades, e.g., lowest C score = $M_{DC} + k\sigma_{DC}$.

Directions to Instructors

a. In each item of the test, cross out, using a single pencil line, those responses which the lowest D-student should be able to reject as incorrect. To the left of the item, against the D-response, write the reciprocal of the number of the remaining responses. (Thus, if you cross out one out of five responses, write 1/4.)

b. Of the remaining responses cross out, using a double line, those which the lowest C-student should be able to reject. Write the reciprocal of the number of responses that still remain to the left of the C-response. (Thus, if you had already crossed out one out of five possible responses, and now cross out two more, write 1/2.)

c. Repeat the procedure for the lowest B-student, using a triple line.

d. Repeat the procedure for the lowest A-student, using a cross.

Example: Light has wave characteristics. Which of the following is the best experimental evidence for this statement?

1 A. Light can be reflected by a mirror.

1 B. Light forms dark and light bands on passing through a small opening.

1/2 C. A beam of white light can be broken into its component colors by a prism.

1/4 D. Light carries energy.

E. Light operates a photoelectric cell.

In the opinion of the instructor who marked the example above, response E should be rejected by the lowest D-student, responses A and D by the lowest C-student, and response C by the lowest B-student. Since the letters of the responses happen to correspond to the usual letter grades, it is convenient to record the reciprocal of the number of responses among which the lowest D-student is

---

to choose against the D-response, etc. In the example above, the lowest B-student is expected to reject all but the correct response; the lowest A-student is of course expected to do just as well; hence number 1 is placed against both response B and response A.

It is possible to construct a test in such a way as to make the determination of the scores corresponding to lowest D, C, B, and A easier and more reliable. In such a test some responses would be designed to be attractive only to F-students, others to F-students and D-students, etc. By including predetermined numbers of such responses the test maker can prepare a test having any desired value for the minimum score corresponding to any letter grade. Whether or not absolute standards are to be used, a test of this kind is likely to have the advantage of being discriminating in the whole range from F to A.

(Nedelsky, 1954, pp.4-10)

Descriptions of the Nedelsky procedure outlined by Glass (1978) and Zieky & Livingston (1977) adapt the original Nedelsky procedure for easier implementation. The Zieky & Livingston description includes a simplified case for only the minimum competence level, while the Glass description includes the consideration of groups of students at different competence levels.

Angoff (1971). In the Angoff (1971) method, expert judges review a test item in its entirety and state the probability that a person with minimum competency can give the correct response. The Angoff procedure is easy to explain, easy to understand, and easy to administer. It is less time consuming than Nedelsky's (1954) and can be used on open-ended items.

In this procedure:

> . . . ask each judge to state the probability that the 'minimally acceptable person' would answer each item correctly. In effect, the judges would think of a number of minimally acceptable persons, instead of only one such person, and would estimate the proportion of minimally acceptable persons who would answer each item correctly. The sum of these probabilities, or proportions, would then represent the minimally acceptable score (Angoff, 1971, p. 515).

Jaeger (1978). In addition, a method proposed by Jaeger (1978), and used for standard setting for student assessment, deserves mention. This procedure maximizes the involvement of educational constituencies. In the North Carolina application, 700 persons convened in groups of 50 to proceed through the standard-setting model. The procedure is as follows:

Judges were first required to take the exam they would later rate. For each item, judges were asked one of the following questions:

1. Should every high school graduate be able to answer this item correctly?
2. If a student does not answer this item, should s/he be denied a high school diploma?

Judges next received the results of the above survey questions as well as actual performance data. With this information, judges were asked to review and revise their initial judgments as they considered necessary.

The procedure then calls for recalculation of the judges' ratings, redistribution of the new ratings, and another judgment. Judges then received information on the proportion of students who would have passed or failed, as determined on the basis of the recommended cutoff scores.

With this information, judges were asked to make a final statement on the "necessity" for each item on the test.

Median scores were calculated by group (type or constituency), and the passing score was then set at the minimum median score calculated for a group.

This process is technically straightforward and involves iterative reviews, and the inclusion of normative student data.

Procedures in Use

Georgia, Alabama. In 1977, Nassif (1978) employed a procedure which began as a modification of Nedelsky. The desire was to simplify the Nedelsky procedure on two dimensions. Each item was to be reviewed in ite entirety, rather than reviewing each component (i.e., each distractor), and one level of competence (minimum acceptable) was considered rather than several. The resulting procedure conceptually matches Angoff. The procedure operationally defined is as follows:

> Panels of expert judges reviewed items independently on an item-by-item basis. The following was asked about each valid item: "Should a person with minimum competency in the teaching field be able to answer this item correctly?" Each judge was asked to imagine the skills of a hypothetical candidate with minimum competency in the content of a teaching field. Within this frame of reference the item was examined as to whether it required too sophisticated a knowledge of the content or whether it required content knowledge of trivial or minor importance.
>
> Judges responded "yes" if the item was considered appropriate for measuring minimum competency or "no" if otherwise. The "I don's know" option was available for judges unfamiliar with the content of an item.
>
> The significance of agreement was determined by comparing the number of "yes" responses with probability tables for the binomial distribution. The ratings of "I don't know" were not considered for any item, so that dichotomous ratings with different numbers of judges were generated. If the probability of receiving a given number of "yes" ratings (i.e., appropriate for minimum competency) was less than a chance of 1 in 10, the item was classified as an appropriate requirement for minimum competency (Nassif, 1978).

This procedure has been used both in the Georgia and Alabama teacher certification programs.

South Carolina, Oklahoma. As mentioned earlier, South Carolina conducted a post hoc validation of the NTE in 1977. In the standard-setting portion of this procedure, modification to the Angoff procedure was used in which judges selected the probability that minimally competence candidates should be able to answer an item correctly from a seven-point scale, rather than providing the probability. While this restricts a judge's choice of response, it eases data reduction and analysis.

In a subsequent teacher certification effort, South Carolina embarked on developing ten content area tests and a basic skills education entrance test. The Angoff procedure as described earlier was used for the content tests. The Jaeger approach was employed for the Basic Skills test.

In the Oklahoma Teacher Certification Program, the Angoff approach was used to determine the standards for the tests.

Florida. The Florida Teacher Certification Exam program involves assessing candidates on competencies in four areas: Math, Reading, Writing, and Professional Education. Each of these areas forms a separate subtest which the candidate must pass. There is, therefore, a separate cut score established for each section. The Writing section is scored holistically and the standard passing score is set by State Board review of performance data and the level of competence described by the score points on the possible performance range.

The cut scores on the three multiple-choice sections are set separately by an Advisory Committee and approved by the State Commissioner. The procedures used to set the cut score involve a review of performance data generated by a field test and an examination of sample items and their associated Rasch calibrations to determine which items represent the cut score.

Advantages

Why are the Nedelsky, Angoff, and Jaeger approaches used predominantly when the list of methods used to set standards for other competency testing programs contains several other standard-setting models? (For example, other methods include: Contrasting Groups and Borderline Groups; Ebel; Administrative Decision (see Nassif, 1979 for a discussion of these models).) Several reasons follow:

- These procedures are based on and permit an item-by-item review. This is a very important consideration for tests that are regenerated in part quite frequently due to test security and job analysis requirements.

- The procedures permit the incorporation of performance data in judgment if desired as additional information in the decision-making process.

- These procedures allow the establishment of single or multiple cut scores as necessitated by the testing program. In the case of multiple cut scores, compensatory or disjunctive scoring can take place.

- These models are easy to understand--a factor which should contribute to the reliability of judges' ratings and to the comprehensibility by constituent audiences.

- These involve and rely on expert judges.

- The cut score that is set does bear a relationship to necessary job performance--a legal requirement. It allows all competent candidates to pass, without restriction from quotas.

- They do not require information (statistical or demographic) not generally available.

- These methods produce a cut score which can be adjusted easily by standard error of measurement to incorporate relevant employment factors.

- These methods can be employed on any number of items, although the original Nedelsky and Jaeger approaches are prohibitive due to the length of the process.

Until recently, few studies had been done comparing the results of using different cut score models. In 1976, Andrew & Hecht found that different cut scores resulted using the Nedelsky and the Ebel procedures. Skakun & Kling (1980) reviewed modified Ebel and Nedelsky procedures, along with their currently used normative approach. While the magnitude of the differences in yielded cut scores varies across comparisons, they found that "results indicate that different approaches for establishing a passing score on an examination produce different standards" (Skakun & Kling, 1980, p. 233). Brennan & Lockwood (1979) found different cut scores produced by Nedelsky and Angoff procedures.

Equating*

Teacher certification testing programs generally provide the candidate with multiple opportunities to retake the exam he/she has failed. If the

---

* The author wishes to acknowledge the contributions to this section of the paper by Dr. Steven Lang-Gunn.

same questions are used repeatedly, the examiner will not know if the candidate's knowledge of the subject matter is being assessed or his/her memory. Another issue that has arisen is that public scrutiny of certification exams may require dissemination of the test even after only a single administration.

> One response to these issues, and perhaps the most prevalent, has been an increased emphasis on development of parallel forms of tests. This response is understandable for three reasons. First, the availability of parallel forms reduces the problem of test security between administrations. Second, it answers political pressure to release the test after administration for use in diagnosis of candidates' weaknesses and tailoring of remedial services. Third, it ensures that an individual student may be retested on the same skills with different test items, minimizing the effects on performance of the prior administration.
>
> The increasing need for alternate forms of tests in programs across the country has redoubled interest among researchers, educators, and policy-makers in how best to ensure that the score or pass-fail decision for a given student not depend on "which form" the student took or "when" the student participated. The statistical problem of test form equivalence takes two primary forms. The first is maximizing the likelihood that a student would receive the same score on two different forms of a test. The second is a more simplified task of minimizing errors of classification--that is, maximizing the likelihood that a student will receive the same classification (pass or fail), although not necessarily the same score, on two alternate forms. The former is most appropriate when the purpose of testing and the prescribed use of test results is to analyze a student's level of functioning and compare it from administration to administration. The latter is most typical of minimum competency testing programs that are directed primarily at determining a student's status simply with respect to a cut score (Nassif, Pinsky, & Rubinstein, 1980).

In practical terms, there are two approaches to accomplishing statistical equivalence of alternate forms. One is to "equate tests" by selecting items with equivalent psychometric characteristics; for example, the p-value method (Nassif, Rubinstein, & Pinsky, 1979) or fit the Rasch model (Wright, 1977). The other is to "equate scores" by paying relatively less attention to the

psychometric characteristics of individual items (except in the normal course of screening for psychometric adequacy) and solving the statistical problem by scaling the test (or subtest) scores produced; for example, the linear and equipercentile methods (Angoff, 1971).

This section is not meant to be a comprehensive technical analysis of equating methods, factors, and consequences. Angoff (1971), Jaeger (1980), Wright (1977), Kolen (1981) to name a few, have presented research on various aspects of this topic. Of primary importance in this discussion is that the purpose is to make the practitioner aware of some aspects of this complex process that so directly affects the area of teacher certification testing. In addition, the reader should know that numerous avenues for guidance or assistance exist for solving these technical issues.

The methods one can use for equating are numerous of course. As in the standard-setting section of this paper, the methods frequently used for teacher certification testing will be described with an indication of which states are adopting which approach.

Following are brief citations of the linear equating technique, the p-value item substitution method of the Rasch model.

Linear Equating (Angoff, 1971). The linear equating model is stated simply as follows. Raw scores are converted to scale scores so that the emphasis is on correct score conversion. Scores are calibrated to adjust for variations in test difficulty and dispersion by using a set of items common to both forms of the test. The purpose of this common section is to establish a statistical link between the two test forms. Through this link, scores on the second form can be calibrated to the scale of the first form. (The approach under consideration here is the one that utilizes a separate test to each group with a common test to both groups (cf. Design IV, Angoff, 1971).

Given two different groups, assumed to be random samples from the same population, as in the case of candidates being tested at various administrations, taking tests x and y, with a common anchor (u) given to both groups (e.g., Group A takes x and u; Group B takes y and u), statistical assumptions are applied to estimate:

$\hat{\bar{x}}_t$ $\quad$ $\hat{s}_{xt}$

$\hat{\bar{y}}_t$ $\quad$ $\hat{s}_{yt}$

for each test, if it were given to the total group (T = A + B).

The goal is to transform raw scores on y (the new form) to the scale of x (the original form). Then, given the estimated parameters, the conversion equation is defined as:

$$x_i = a y_i + 6$$

Where $a = \frac{s_{xt}}{s_{yt}}$ and $b = \hat{\bar{x}}_t - a\,\hat{\bar{y}}_t$

The Tucker model of linear equating is used when the two groups do not differ widely in ability as measured by performance on u. The Levine method is used when the two groups do differ.

Forms of the National Teachers Examination, administered several times a year, are equated (Angoff, 1971) by linear or equipercentile methods.

The Alabama and Oklahoma State Teacher Certification Testing Programs are designed to use the Tucker linear equating method. The anchor tests in these programs are the subset of items which are repeated across two successive administrations, i.e., common to both administrations. The anchor tests

contain the same distribution of items by content area as each total test and comprise 70-80% of each total test. New items replace previously used items matched on content and difficulty. Upon analyzing the test data, items with the best statistical properties representing the strongest content coverage constitute the scorable items on the new main form. The scorable items are equated to the same number of scorable items on the form previously administered. The data are reported on a converted scale which allows the same reported cut score across different test forms or fields and multiple administrations.

The advantage of using linear equating is that equivalence of scoring is ensured. However, a disadvantage in the teacher certification environment occurs in teaching fields with too low incidence of candidate examinees for equating. In low incidence fields, the p-value approach, described later, may be more appropriate. An advantage to the linear equating method is that items need not be field tested prior to the administration in which they are used as scorable items.

It should be noted that linear equating is only appropriate when the relationship between the raw scores and the transformed scores is, in fact, linear. Where significant deviations from linearity are observed, equipercentile methods should be used (Jaeger, 1980).

P-value/Point-Biserial Test Equating. This straightforward plan requires the construction of tests with equating, replaceable, and experimental sections. Each of these sections is a mini-test, in that it forms a stratified sample of items from the entire test domain. (The pass/fail decision is based on the scorable replaceable items; that is, the experimental items do not contribute to the examinee's score.) The substitution of experimental items into scorable replaceable items is done within an objective

for items of the same difficulty and comparable point-biserial (item-total test discrimination) as determined from the previous testing session. The item substitution plan preserves the content validity of the test as well as the statistical difficulty of the test. Where items cannot be matched exactly on p-value within an objective, one averages the differences over clusters of objectives within the same content subarea (Nassif, Pinsky, & Rubinstein, 1979).

This method has been used successfully in the Georgia Teacher Certification Program.

Rasch Model. According to Wright (1968), the Rasch model calibrates test items independent of the ability level of the examinee sample used for calibration purposes. Further, the measurement of examinees occurs independent of the difficulty of the test it has used for measurement purposes.

Since sample-free estimates of item difficulty with respect to a common score are obtained for all items, item banking is easily achieved. Parallel forms of tests are then created and equivalence of scoring is ensured by creation of test forms of known difficulty and dispersion.

The Florida and parts of the South Carolina Teacher Certification Testing Programs rely on the Rasch model for creating equated tests used in successive administrations. Items from previous administrations are seeded onto subsequent test forms to observe shifts. These seeded items also provide a link back to the item bank.

Why Are These Methods Used?

- Linear equating is a straightforward procedure which accommodates varying amounts of item overlap from one administration to the next. Generally, it is advised that at least 25% of the test be anchored from one administration to the next.

- Different linear equating methods available accommodate varying statistical assumptions or effects (e.g., Tucker & Levine).

- Linear equating does not require a separate field test of the new replacement items, assuming sufficient sample size.

- All models allow for, but may not require, content mapping and difficulty and discrimination match of the replacement with the replaceable items.

- The p-value approach for creating new forms accommodates teaching fields with low incidence of applicants in that data are pooled over several administrations until an adequate data base has accumulated.

- The Rasch model targets tests to examinees' ability level, so that greater efficiency in testing is believed to be achieved.

When Do Test Forms Need to be Changed?

If there is reason to believe that there has been a security break on the test, a new test form should be developed and administered. After a test form has been administered several times and there is reason to believe that the performance on the test can be significantly affected by multiple retakes

of the same exam, a new exam should be developed. Clearly, if the content domain or job definition changes, corresponding changes should be reflected on the test.

In this time of restricted resources. no test administrator wants to develop more tests than are necessary. In the teaching fields with few examinees, multiple administrations of the same exam can be justified for the purpose of test statistical data collection. In larger fields, the test form may be changed after it has been administered to an adequate number of examinees (say, 250). This generally occurs at least once a year in these larger fields.

Summary

Technical apsects in teacher certification program design need careful attention. Several states, notably Georgia, Florida, South Carolina, Oklahoma, and Alabama have begun addressing these issues of validity, job analysis, standard setting and equating, and embarked on various developmental efforts. Other states are in the process of examining these very issues. Their solutions will be viewed with much interest. Many resources are available to the administrator/policy/decision makers thrust into addressing matters of legal and technical composition and consequence. The field is replete with the need for further developmental efforts in these issues and the corresponding talent and interest to satisfy those needs.

References

Andrew, B. J., & Hecht, J. T. A preliminary investigation of two procedures for setting examination standards. Educational and Psychological Measurement, 1976, 36, 45-50.

Angoff, W. H. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Baker v. Columbus Municipal Separate School District, 329 F. Supp. 706 (1971).

Brennan, R. L., & Lockwood, R. E. A comparison of two cutting score procedures using generalizability theory. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, April, 1979.

Ebel, R. L. Essentials of educational measurement. Englewood Cliffs, New Jersey: Prentice-Hall, 1972.

Georgia Association of Educators v. Jack P. Nix, 407 F Supp. 1102 (1976).

Glass, G. Standards and criteria. Journal of Educational Measurement, 1978, 15, 237-261.

Hambleton, R. K., & Eignor, D. R. Competency test development, validation, and standard-setting (Research Rep. No. 84, Laboratory of Psychometric and Evaluation Research). Amherst, Massachusetts: University of Massachusetts, School of Education, 1978.

Jaeger, R. M. A proposal for setting a standard on the North Carolina High School Competency Test. Paper presented at the meeting of the North Carolina Association for Research in Education, Chapel Hill, North Carolina, 1978.

Jaeger, R. M. Some exploratory indices for selection of a test equating method. Paper presented at the Annual Meeting of the American Educational Research Association, Boston, April, 1980.

Koffler, S. L. A comparison of approaches for setting proficiency standards. Journal of Educational Measurement, 1980, 17 (3), pp. 167-178.

Kolen, M. J. Comparison of traditional and item response theory methods for equating tests. Journal of Educational Measurement, 1981, 18 (1), 1-11.

Nassif, P. M. Standard-setting for criterion-referenced teacher licensing tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Toronto, March, 1978.

Nassif, P. M. Setting standards. In Final Program Development Resource Document: A Study of Minimum Competency Testing Programs, National Institute of Education, December, 1979.

Nassif, P. M., Pinsky, P. D., & Rubinstein, S. A. Generating parallel test forms for minimum competency exams. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, April, 1979.

Nassif, P. M., Pinsky, P. D., & Rubinstein, S. A. Further work developing parallel tests by p-value substitution. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Boston, April, 1980.

Nedelsky, L. Absolute grading standards for objective tests. Educational and Psychological Measurement, 1954, 14, 3-19.

Skakun, E. N. & Kling, S. Comparability of methods for setting standards. Journal of Educational Measurement, 1980, 17, 229-235.

United States v. North Carolina, 400 F Supp. 343 (E.D.N.C. 1975), 425 F. Supp. 789 (E.D.S.C 1977).

Wright, B. H. Sample-free test calibration and person measurement. Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, NJ, Educational Testing Service, 1968.

Wright, B. H. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116.

Zieky, M. J. & Livingston, S. A. Manual for setting standards on the basic skills assessment tests. Princeton, NJ: Educational Testing Service, 1977.

TEACHER CERTIFICATION TESTING

TECHNICAL CHALLENGES: PART II

Scott M. Elliot

National Evaluation Systems, Inc.
30 Gatehouse Road
Amherst, MA 01004

A Paper Presented at the Annual Meeting
of The National Council on Measurement
in Education, New York, 1982

Job Analysis

Any instrument designed for certification or licensing, as is the case in teacher certification testing, must be shown to be job related. It must fairly measure the content knowledge relevant to the job as performed by present job incumbents. Determining the job relatedness of content selected for inclusion in certification tests is both endorsed in the APA Principles for the Validation and Use of Personnel Selection Procedures (1980) and required by the Equal Employment Opportunity Commission Guidelines (1978). The guidelines require that the criteria used as a basis of certification must bear an empirical and logical relationship to successful job performance. For purposes of teacher certification, this suggests that test content should reflect the content knowledge or pedagogical skills required for teaching. While there are a number of ways in which this domain of knowledge can be identified (cf. Popham, 1980), a systematic job analysis is recommended to establish an empirical and logical relationship to teacher performance.

Job Analysis Approaches

Job analysis is a process of systematically collecting information about the elements of a job. While job analysis has been routinely used in personnel-related areas for close to a century, it is only within the past few decades that it has been employed in personnel testing.

A variety of approaches to assessing the elements of a given work

situation are available; however, regardless of the selected method, most approaches include some determination of the critical and frequently performed elements of the job. Importance (criticality or essentiality) and frequency of performance (time spent or percentage of time consumed on job) are the two key dimensions underlying most job analysis approaches. Within the teacher certification arena, this would generally take the form of assessing the important and frequently applied teaching skills or content knowledge in the instructional setting.

Job analysis approaches can be seen to vary along a number of dimensions. Levine, Ash, Hall, and Sistrunk (1981) have delineated three key dimensions along which job analyses vary:

- type of descriptor or element used to describe the job,
- the source of job information, and
- data collection methodology.

Among the descriptors used to describe a job are tasks, activities, skills, knowledge, and personal characteristics. A number of sources of job information are potentially available; these include job incumbents, supervisors, trained job analysts, and written documents. Data collection methods include questionnaires, interviews, observation, diaries, and actual job performance. Although it is clear that many approaches to conducting job analyses are available, the application of job analysis methodology to teacher certification testing has been somewhat limited.

Current applications of job analysis methodology to teacher certification testing is presented below. Other job analysis approaches derived from the three dimensions cited previously, with potential applications to teacher certification, are offered following the discussion of current applications.

Job Analysis Applications

Job analysis has been used in the content validation of teacher certification tests in a number of states. Among the states that have conducted job analyses as part of their teacher certification test development efforts are Georgia, Alabama, South Carolina, and Oklahoma. In all four cases a survey approach was used. A sample of educators within the state were sent a survey instrument requesting them to rate on a Likert-type scale a series of content objectives, developed by panels of content experts, in terms of the amount of time spent teaching or using the objectives and the extent to which the objectives were essential to the field. Based on the job analysis results, those objectives found to be most job related were included in the content of the examinations. In some cases an interview procedure was used with a sample of educators to supplement the quantitative ratings and gather further information about job content.

Similar procedures were used in the development of the Florida Teacher Certification Examination. Teacher competencies (objectives) were developed by a panel of teacher educators. The competencies were then sent to a sample of educators who rated the competencies in terms of their perceived "importance" to the field. No ratings of "frequency of use" or "time spent using" were collected.

Similar procedures have been used for more process-oriented assessment measures developed for use in teacher certification. The Basic Professional Studies Examination developed in Alabama to assess knowledge of pedagogical skills relied on job analysis for determining the content to appear on the test. A sample of educators across teaching fields rated the frequency with which pedagogical skills were used and the importance of those skills. The content of the Performance Observation Instrument developed in South Carolina was defined through job analysis procedure. Again, using a survey approach, a sample of South Carolina educators rated the importance and frequency of use (as well as observability and relevance) of a series of teaching skills and behaviors.

The development of the current NTE did not involve job analysis; however, the "Common" portion of the NTE is currently under revision, and a form of job analysis is being used in defining the content to be included on the revised examination. Here, state representatives have been surveyed to determine the extent to which a proposed set of pedagogically related topics are important for purposes of teacher certification.

Job Analysis Alternatives

While job analyses conducted for current teacher certification tests have almost exclusively been limited to survey questionnaires requesting job incumbents to rate proposed test content in terms of importance and

frequency of use, other alternatives suitable for use in teacher certification testing are available. Recommendations for job analysis alternatives based on (1) type of data collection methodology and (2) source of job information are provided below.

Teacher certification test development efforts, to date, have relied on the collection of job information from a cross section of job incumbents reflecting the teaching area for which the measure is being developed. Alternatives to the use of a cross section of job incumbents include the collection of job information from supervisors or solely from superior performers on the job. Previous research comparing the job information obtained from job incumbents and other observers is conflicting (Levine et al., 1981). While the information obtained from incumbents and other observers appears to be consistent in some job settings, Levine et al., (1981) suggest that in other settings incumbents tend to provide less accurate accounts of their job content. No specific attempts have been made to investigate the information obtained from teachers as compared to the information obtained from other observers in the instructional environment, and the accuracy of teacher/educator supplied information remains to be explored. Future job analysis efforts within the realm of teacher certification should consider obtaining information from teacher supervisors (or outside observers) as well as from teachers for purposes of comparison.

Similarly, little effort has been made to compare the job information obtained from teachers judged as superior to educators judged to be poor performers. While Levine et al. (1981), in their recent discussion of job analysis methodology, suggest that there are few differences in the job

information obtained from superior and less capable performers in a variety of job settings, this remains to be verified in the instructional setting. Future efforts to determine job-related content for inclusion in teacher certification assessment instruments should include the examination of the differences in information provided by educators exhibiting different levels of performance (previously identified by school personnel). However, the validity of teacher certification tests based on job content defined solely by superior performers could be brought into question as these measures are generally designed as minimum competency assessments.

Alternative data collection methods should be considered in job analysis efforts undertaken for teacher certification test development purposes. Among the alternatives to the survey questionnaire approach (which has been the primary data collection method employed for teacher certification testing to date) are (1) observation, (2) critical incident technique (Flanagan, 1954), (3) document review, and (4) group discussion.

Job analysis data collection using observational methods relies on trained observers observing the performance of job incumbents. Within the realm of teacher certification, this would involve trained observers observing the classroom behavior of teachers or other instructional personnel to ascertain the content of the job. While providing a direct assessment of the job content, the feasibility of this approach is questionable because of its obtrusiveness and resources required. This is particularly true in the case of content knowledge examinations developed for teacher certification purposes; repeated observations over an extended

period of time would be required to provide an accurate assessment of the content knowledge required on the job.

The critical incident approach, developed by Flanagan (1954), involves the identification of job events that have resulted in either inferior or superior performance (i.e., events that elicit behaviors necessary for successful job performance). A large number of incidents are collected from job incumbents (through diaries, interviews, etc.) and are used to determine what behaviors are necessary to be effective on the job. In application to teacher certification testing, this would require the elicitation of critical incidents in the instructional setting from a pool of instructional personnel. This approach is potentially useful for the development of teacher performance measures or tests focusing on pedagogical skills; however, the critical incident technique appears to have little application to content knowledge-oriented measures. Levine et al., (1981) report that this approach was not favored by experienced job analysts for use in personnel selection.

The final two data collection approaches with potential application to teacher certification are document review and group discussion. Document review involves the use of available literature defining a job as a basis for determining necessary job content. Here, job descriptions and other documentation would be reviewed to determine the critical aspects of the job to include in a personnel selection instrument. To the extent that such documents exist within educational environments, this approach could be employed. In fact, the review of such documents is already carried out, to a limited extent, in the definition of content knowledge or skills to be included on job analysis survey instruments used in existing teacher certification test development projects. Similarly, the

fourth and final method to be considered--group discussion--has been used in the development of existing teacher certification tests. In the development of certification tests for Georgia, South Carolina, Alabama, and Oklahoma, panels of experts were convened in the respective content areas to generate content for inclusion on the job analysis survey instrument. This could be expanded to include supervisors and incumbents in the respective areas who would formally rate the knowledges and skills identified in terms of their importance as is recommended by Primoff (1975).

Whether the additional information gained from the use of these approaches warrants the large expenditure of resources remains to be seen. However, additional research in this area is necessary to determine the effectiveness of current job analysis approaches employed within the realm of teacher certification, and to identify superior approaches to job analysis in this setting.

Validity

One of the primary concerns in the teacher certification measurement effort is validity. Validity refers to the ability of a measuring instrument to do what it is intended to do (Nunnally, 1978), or, more specifically, "the degree to which inferences from scores on tests or assessments are justified or supported by evidence" (APA Principles, 1980, p. 2). Traditionally, and in licensing, three aspects of validity are discussed: criterion-related validity (predictive and concurrent), content validity, and construct validity (APA Standards, 1974). Criterion-related validity is of concern when one wishes to infer, from a given instrument, an individual's performance on some other variable referred to as the criterion (APA Standards, 1974; Nunnally, 1978). Content validity is of importance when one wishes to estimate "how an individual performs in the universe of situations the test is intended to represent" (APA Standards, 1974). The third aspect of validity, construct validity, references the extent to which a measurement tool is related to the various elements or underlying traits associated with the psychological construct it is purported to measure.

Validity is of particular concern in the development and use of personnel screening instruments where one wishes to establish that a test does indeed truly measure the important aspects of job performance it is purported to measure. It is imperative that a relationship between teacher certification decisions based on a measurement instrument and aspects of the job required for successful performance in the classroom be established. Most of the validation attempts for teacher certification

tests have focused on content validity. The key concern within teacher certification testing has been to ensure that the tests developed reflect the significant aspects of the teaching profession for which they are designed. At a minimum, the content of certification instruments should be drawn from important elements of the teaching job.

A discussion of the validation of teacher certification assessment measures is provided below. As most of the validation in teacher certification tests is focused on content validation, the focus of the discussion provided is on content validity.

Content Validity

The content validity of a test is established by demonstrating that the content included within the instrument represents a sample of the content or behavior included in the performance domain. Content validation, as applied to teacher certification, generally has two components: (1) determining whether the test content reflects significant aspects of the educator's job (and measures those aspects proportionally), and (2) determining whether test items developed accurately measure that job's content. The first component is often assessed through some form of job analysis and is discussed at length in earlier sections of this paper. Discussion of the second area, item validation, is presented in the following sections.

Content Validation Approaches. A variety of approaches to assessing item validity are available to the practitioner. Among the methods avail-

able to be considered here are (a) index of item-objective congruence, (b) rating scale approach, and (c) dichotomous judgment model. Within each of the above approaches, a panel of judges evaluates examination items on an item-by-item basis to determine if the item is a valid measure of the domain (objectives, item specification, topic) for which it was written.

Within the item-objective congruence model, content experts are asked to assign ratings of +1 (item measures the objective), 0 (undecided whether item measures the objective) and -1 (item does not measure the objective) to each item. Judges are asked to rate each item against each objective. An index of item-objective congruence (ranging from 1 to -1), developed by Rovinelli and Hambleton (1977), can then be computed for each item, and a cutoff score for identifying items as valid or invalid can be established.

The rating scale approach (Hambleton, 1980) involves expert judges assessing each item as a measure of its intended objective, on a rating scale. The mean or median score across judges is computed and a cutoff score for accepting items as valid is set. The index of item-objective congruence and rating scale procedures are described in more depth in Hambleton (1980).

A third approach available is the dichotomous judgment model (Nassif, 1978). Here a panel of content experts indicate, for each item, whether they feel the item is or is not a valid measure of the objective for which it was written. Item validity is defined as having four parts: accuracy, congruence with objective, significance, and lack of bias. The

results from the content expert evaluations for each item are compared to the binomial distribution to determine the probability, due to chance alone, of obtaining "x" valid responses for an item from a total of "N" raters. Items receiving ratings meeting statistical significance are treated as valid.

Content Validation Applications. Content validation procedures have been employed in a variety of teacher certification testing efforts. The dichotomous judgment model has been widely used in the validation of content knowledge examinations in a number of states. This approach has been used in the development of certification tests for educators in Georgia, where panels of approximately 15 content experts in each field for which examinations were developed were asked to make dichotomous judgments about prospective items to be included on the test. For each item where the probability of obtaining "x" valid responses fron "N" raters due to chance alone was less than .10, the item was categorized as valid. Similar procedures were employed in the development of content examinations for teacher certification in Alabama and Oklahoma. The dichotomous judgment model was also applied in the development of the English Language Proficiency Examination to be administered to individuals seeking admission to teacher education programs in Alabama.

The content validation of the teacher certification tests developed for Florida involved the review of test items by two independent panels of experts in the four subtest areas. The panels reviewed the items based on supplied criteria (e.g., item-competency match, bias) and recommended acceptance, rejection, or revision of each item.

Post hoc efforts to establish the validity of the National Teacher's Examination (NTE) have been undertaken in a number of states. In South Carolina, a validation study was undertaken to, among other reasons, establish the content validity of the NTE. Panels of content experts in the various NTE teaching areas were asked to judge whether the individual questions appearing on the examination were covered in the curriculum of South Carolina teacher education programs. This is akin to the dichotomous judgment model presented earlier; however, 51% or more of the judges citing the item as congruent with the curriculum was employed as the criterion for accepting items as valid, rather than relying on comparisons to the binomial distribution. Similar validation efforts for the NTE are planned or are underway in Arkansas, Kentucky, Virginia, and Tennessee.

Item validation for the South Carolina Teaching Area Examinations, administered to teachers exiting teacher education programs for purposes of initial certification, relied on the rating scale approach. Panels of South Carolina educators rated each item developed on a scale from 1 to 5 ranging from "clearly valid" to "clearly not valid." Items receiving mean ratings, below 3.0 across judges, were treated as valid.

There has been little attempt to apply item-objective congruence models in teacher certification to date. The primary reason for its absence from the teacher certification area stems from issues of feasibility. The approach is quite time-consuming and potentially quite costly to the consumer. For example, if there are 50 objectives and 100 test items, each judge must make 5,000 judgments.

The rating scale approach and dichotomous judgment model offer practical advantages; they are relatively simple to administer, and the analysis associated with them is fairly straightforward. The dichotomous judgment model, when used in conjunction with the binomial distribution, offers the added advantage of preventing the assessment of items determined to be valid based on chance alone.

While content validity is clearly an important element in the development of teacher certification tests, a number of measurement specialists have emphasized that content validity is an insufficient criterion for establishing the validity of a test. Messick (1975) and, more recently, Hambleton (1980) note that content validity does not provide evidence regarding the uses of or inferences made from test scores. Despite the importance assigned to criterion-related validity and construct validity, few validation studies in this area have been conducted in the teacher certification field. These issues are discussed in greater depth in the following sections.

Criterion-Related Validity

Criterion-related validity "compares test scores or predictions made from them, with an external variable (criterion) considered to provide a direct measure of the characteristic or behavior in question" (Cronbach, 1971, p. 444). Criterion-related validity, as applied to teacher certification, examines the relationship between an instrument administered for

certification purposes and actual teacher performance on the job. A teacher certification test should accurately predict that aspect of teacher competency for which it was designed.

Two forms of criterion-related validation are generally discussed: (1) concurrent validity, and (2) predictive validity (APA Standards, 1974). "Statements of concurrent validity indicate the extent to which the test may be used to estimate an individual's present standing on the criterion," whereas predictive validity refers to "the extent to which an individual's future level on a criterion can be predicted from a knowledge of prior test performance" (APA Standards, 1974, p. 26). Concurrent validation, as applied to teacher certification testing, examines the relationship between the test scores of practicing educators (job incumbents) and current performance. Establishing the predictive validity of a teacher certification measure involves the examination of the relationsnip between the test scores of prospective teachers (job applicants) and future performance. Both forms of criterion-related validity are concerned with the accuracy of the measures in predicting teacher competency.

While criterion-related validity has been held as a necessary part of validating certification tests, a number of obstacles have prevented the execution of criterion-related validation studies for teacher certification measures. Hecht (1976), while supporting the importance of criterion-related validation for licensing and certification tests, notes that criterion-related validation studies are "difficult to develop, time-consuming, impractical for numerous reasons, and expensive" (p. 8). Nassif, Gorth, and Rubinstein (1977) provide a more in-depth treatment of these issues, as they relate specifically to teacher certification

testing. Nassif et al. (1977) suggest that in order to demonstrate the predictive validity of teacher certification tests, the following criteria are required:

(1) admission of all applicants for employment in the field;

(2) sufficient time lapse before observing the criterion variable;

(3) unexamined, unused results of the test, i.e., the predictor stored until correlated with the criterion (here, retention or dismissal of teacher due to subject-matter competence/ incompetence);

(4) the criterion must be measurable, i.e., a mechanism for accurately and reliably collecting the reasons for retention or dismissal of teachers (criterion) which clearly separates content knowledge as one of those reasons;

(5) sufficient sample size; and

(6) stability of the criterion.

However, these factors are usually not present in a certification program. Problems associated with conducting a criterion-related validation study for teacher certification tests are discussed at length by Nassif et al. (1977).

Construct Validity

Construct validity is aimed at answering the question "Does the test measure the attribute it is said to measure" (Cronbach, 1971). Construct validation is a process (rather than a single study) whereby evidence, relating test scores to the attributes of the construct the test is purported to measure, is accumulated. Cronbach (1971) notes that when statements are made that test scores reflect levels of a certain skill or knowledge, one is "constructing" an interpretation of these scores, and construct validation is of necessity. While the constructs underlying teacher certification measures are somewhat simpler than those encountered in a more complex and abstract personality construct such as "aggressiveness," there exists an underlying construct (i.e., "pedagogical skill" or "content knowledge").

The construct validation of tests designed for teacher certification presents a number of problems, and, as such, there has been little effort to construct validate existing teacher certification measures. Potential approaches to, and problems inherent in, conducting construct validation studies in this area and the inherent problems in conducting construct validation studies in this area are discussed below.

One of the primary methods for establishing the construct validity of a given measure is to establish a relationship between that measure and other measures of the same construct. For content knowledge tests used for teacher certification purposes, this would require a comparison of the

tests with other assessments of applicants' content knowledge. Similarly, performance or pedagogical skill certification tests would be compared with alternative performance or pedagogical skill assessments. Attempts to construct validate teacher certification tests using alternative measures of the construct suffer from many of the problems noted earlier in our discussion of criterion-related validity, notably the location of a suitable criterion measure and the stability of that criterion. A "well-matched" criterion measure adequately measuring the construct reflected in the test to be validated is often unavailable. Moreover, the use of instructor or supervisor assessments of a candidate's proficiency are unsuitable as criterion measures for construct validation because of the unreliability and questionable accuracy of such criteria.

While it is difficult to obtain suitable criterion measures for use in the construct validation of teacher certification tests, Hambleton (1980) notes that construct validation should also be aimed at examining possible sources of error that reduce the validity of test scores. Among the factors suggested for consideration by Hambleton (1980) applicable to teacher certification are the effects of test administration procedures, examinee test taking skills, and examinee motivation. Although little attempt has been made to investigate the impact of these factors on teacher certification, future validation efforts in this area should include the consideration of these factors. Another approach to construct validation, suggested by Hambleton (1980), involves the use of factor analysis to verify the domain structure of the test. One would expect the factor structure of the test to correspond to the domain structure of the

test design, with individual test items loading on a single factor corresponding to the appropriate domain. This approach has been employed in the development of the teacher performance assessment instruments in Georgia and South Carolina.

Reliability

Reliability concerns the extent to which a measure consistently produces the same result under similar conditions (Nunnally, 1978). As with any measurement effort, the reliability of assessment instruments used is a key concern in teacher certification tests. Traditionally, reliability has been thought of in terms of the internal consistency of a test or the stability of test scores across repeated administrations and parallel forms of the test. More recently, particularly in the area of certification, test developers have begun to examine reliability in terms of the dependability of classification decisions (e.g., pass/fail). Traditional and more recent approaches to teacher certification test reliability, and their current applications, are considered below.

Approaches

A number of methods for determining the reliability of teacher certification tests are available. Traditionally, three approaches to reliability have been employed: (1) stability, (2) equivalence, and (3) internal consistency. Stability refers to the consistency of the measurement

over time, while equivalence estimates are obtained to determine the consistency of measurement across two or more forms of the test. The internal consistency of a test refers to the consistency of items included within a single test form. By far the most common approach is internal consistency estimates because of the need for only one test form and the ease with which these estimates can be obtained.

The most common approach to assessing the stability of a test over time is the test-retest method, where the same test is administered to a single group of individuals at two different points in time. The correlation between the scores at T_1 and T_2 is obtained as an estimate of the test's reliability (Nunnally, 1978). Similarly, the reliability of two alternative forms (equivalence) can be determined by administering two forms of a test to a pool of examinees and computing the correlation between the two sets of scores as an estimate of test reliability (Nunnally, 1978). However, this approach has little application in teacher certification testing, as only a single test form is employed in most certification programs.

Internal consistency approaches estimate test reliability using a single test form. Two internal consistency approaches are generally employed: split-half reliability and the Kuder-Richardson indices of item homogeneity (K-R20, K-R21; Nunnally, 1978). The former approach involves the splitting of a test into two halves and correlating the two sets of items as an estimate of internal reliability. The latter approach examines the average of all possible split-half reliability coefficients. While both internal consistency approaches are used, the Kuder-Richardson formulas are considered more accurate and hence are employed with greater frequency.

More recently, a number of writers (cf. Huynh, 1976) have suggested that the reliability of tests, in situations where a dichotomous decision is made on the basis of test scores, should be assessed on the basis of the consistency of the decisions across test administrations. It has been suggested that this is particularly applicable in criterion-referenced testing where the problem of restricted range of test scores may be present. These approaches would seem particularly applicable to the teacher certification testing area where dichotomous master/nonmaster decisions are made.

While a number of decision-consistency approaches have emerged in recent years, only a sample of the more visible approaches applicable to teacher certification are presented here. Among the available approaches discussed here are Kappa reliability (Swaminathan, Hambleton, and Algina, 1974; Huynh, 1976; Subkoviak, 1980) and generalizability analysis (Brennan, 1980).

The Kappa reliability approach examines the consistency of classification decisions across test administrations. The extent of actual agreement across test administrations (computed by calculating the proportion of examinees consistently classified in a given mastery state on two administrations) is compared to the extent of agreement that could be expected by chance alone. These two facets are used to calculate a coefficient of decision-consistency. Specific procedures for computing Kappa are described in Swaminathan et al., 1974. Procedures for obtaining Kappa reliability estimates from a single test administration are discussed in Huynh (1976) and Subkoviak (1980). The assessment of reliability using generalizability theory employs estimates of the variance components

attributable to the various elements in the assessment situation (e.g., items, persons). Reliability, then, is viewed as a function of the proportion of variance accounted for by the person component.

Applications

While there are a considerable number of approaches to examining test reliability, the diversity of applications of test reliability methods to teacher certification testing has been somewhat limited.

The traditional approaches to reliability have been extensively applied in teacher certification, particularly internal consistency approaches. For virtually all teacher certification measurement efforts, internal reliability estimates have been obtained. K-R20 reliability coefficients are routinely obtained for teacher certification tests administered in Georgia, Alabama, Oklahoma, and other statewide certification programs, as well as for the NTE. This is not surprising as these estimates are reasonably easy to obtain and provide a reasonable assessment of test reliability.

With increased criticism of more traditional reliability approaches, test developers in the area of teacher certification have begun to employ decision-consistency models and generalizability analysis with increased frequency. The reliability of the teacher performance assessment instrument in Georgia was recently examined using generalizability analysis. Both decision-consistency (Subkoviak, 1980) and generalizability analysis were applied in the assessment of the reliability of the Georgia teaching field examinations.

References

American Psychological Association, Division of Industrial-Organizational Psychology, Principles for the Validation and Use of Personnel Selection Procedures: Second Edition, Berkeley, CA: Author, 1980.

Brennan, R. L. Applications of Generalizability Theory in R. A. Berk (Ed.) Criterion-referenced testing: The state of the art, Baltimore, MD: John's Hopkins University Press, 1980.

Cronbach, L.J. "Test Validation" in R.L. Thorndike (Ed.) Educational Measurement, Washington, D.C.: American Council on Education, 1971.

Flanagan, J.C. "The Critical Incident Technique," Psychological Bulletin, 1954, 51, 327-358.

Hambleton, R.K. "Test Score Validity and Standard-Setting Methods," in R.A. Berk (Ed.) Criterion referenced measurement: The state of the art, Baltimore, MD: Johns Hopkins University Press, 1980.

Hecht, K.A. "Professional Licensing and Certification: Current Status and Methodological Problems of Validation," paper presented at the annual convention of NCME, San Francisco, 1976.

Huynh, H. On the reliability of decisions in domain referenced testing, Journal of Educational Measurement, 13, 256-264, 1976.

Levine, E.L., Ash, R.A., Hall, H.L. and Sistrunk, F. "Evaluation of Seven Job Analysis Methods by Experienced Job Analysts," unpublished research report, Center for Evaluation Research, University of South Florida, 1981.

Messick, S.H. "The standard problem: Meaning and values in measurement and evaluation," American Psychologist, 30: 955-66, 1975.

Nassif, P. M. Standard-Setting for Criterion-Referenced Teacher Licensing Tests, paper presented at the annual meeting of the National Council on Measurement in Education, Toronto, March, 1978.

Nassif, P.M., Gorth, W.P., and Rubinstein, S.A. "Developing and Validating Teacher Certification Tests According to Federal Guidelines," 1977.

Nunnally, J.C. Psychometric Theory, New York: McGraw-Hill Publishing Co., 1978.

Popham, W.J. "Domain Specification Strategies" in R.A. Berk (Ed.) Criterion-referenced testing: The state of the art, Baltimore, MD: Johns Hopkins University Press, 1980.

Primoff, E.S. "How to Prepare and Conduct Job Element Examinations," Washington, D.C.: U.S. Government Printing Office, 1975.

Rovinelli, R.J. and Hambleton, R.K. "On the Use of Content Specialists in the Assessment of Criterion-Referenced Test Item Validity," Dutch Journal of Educational Research, 2:49-60, 1977.

Subkoviak, M. Decision-consistency approaches in R. A. Berk (Ed.), Criterion-referenced testing: The state of the art, Baltimore, MD: Johns Hopkins University Press, 1980.

Swaminathan, H., Hambleton, R. K. and Algina, J. Reliability of criterion-referenced tests: A decision theoretic formulation, Journal of Educational Measurement, 11, 263-267, 1974.